Journal of Science Education and Technology (2020) 29:136–160
https://doi.org/10.1007/s10956-020-09811-1

Modeling and Measuring High School Students' Computational Thinking Practices in Science

Golnaz Arastoopour Irgens[1] · Sugat Dabholkar[2] · Connor Bain[2] · Philip Woods[2] · Kevin Hall[2] · Hillary Swanson[2] · Michael Horn[2] · Uri Wilensky[2]

Published online: 5 February 2020

Abstract

Despite STEM education communities recognizing the importance of integrating computational thinking (CT) into high school curricula, computation still remains a separate area of study in K-12 contexts. In addition, much of the research on CT has focused on creating generally agreed-upon definitions and curricula, but few studies have empirically tested assessments or used contemporary learning sciences methods to do so. In this paper, we outline the implementation of an assessment approach for a 10-day high school biology unit with computational thinking activities that examines student pre-post responses as well as responses to embedded assessments throughout the unit. Using pre-post scores, we identified students with both positive and negative gains and examined how each group's CT practices developed as they engaged with the curricular unit. Our results show that (1) students exhibited science and computational learning gains after engaging with a science unit with computational models and (2) that the use of embedded assessments and discourse analytics tools reveals how students think differently with computational tools throughout the unit.

Keywords Computational thinking · Learning analytics · Assessment · Biology · Science learning

Introduction

In recent decades, computational tools and methods have become pervasive in mathematical and scientific fields (National Research Council 2010). Tools such as mathematical and statistical models have expanded the range of phenomena that are explored and have become necessary for analyzing increasingly large data sets across disciplines (National Academy of Sciences, National Academy of Engineering,, and Institute of Medicine 2007). With these advances, entirely new fields such as computational statistics, neuroinformatics, and chemometrics have emerged. The varied applied uses of computational tools across these fields have shown that future scientists will not only need to know how to program but also be knowledgeable about how information is stored and managed, the possibilities and limitations of computational simulations, and how to choose, use, and make sense of modeling tools (Foster 2006).

As a result of these changes, science, technology, engineering, and mathematics (STEM) education communities have recognized the importance of integrating computational thinking (CT) into school curricula (National Research Council 2012; NGSS Lead States 2013), and there are several important efforts underway to more closely integrate CT skills and

✉ Golnaz Arastoopour Irgens
garasto@clemson.edu

Sugat Dabholkar
SugatDabholkar2020@u.northwestern.edu

Connor Bain
connorbain2015@u.northwestern.edu

Philip Woods
philipwoods93@gmail.com

Kevin Hall
kevin.hall.1@northwestern.edu

Hillary Swanson
hillary.swanson@northwestern.edu

Michael Horn
michael-horn@northwestern.edu

Uri Wilensky
uri@northwestern.edu

[1] Clemson University, Clemson, SC, USA

[2] Northwestern University, Evanston, IL, USA

practices into mainstream science and mathematics classrooms such as Bootstrap (Schanzer et al. 2018; https://www.bootstrapworld.org), GUTS (Lee et al. 2011; https://teacherswithguts.org), and CT-STEM (Swanson et al. 2019; https://ct-stem.northwestern.edu). However, while much of the research on CT and CT in STEM has focused on creating generally agreed-upon definitions and CT curricula (Shute et al. 2017), few studies have empirically tested assessments or used contemporary learning sciences methods to do so (Grover and Pea 2013). In this paper, we outline the assessment approach for a 10-day biology unit with computational thinking activities. We examine both high school student pre-post responses as well as responses to embedded assessments throughout the unit. We explain how we coded responses for CT-STEM discourse elements and then quantitatively measured the development of students' CT-STEM practices over time. We identify two groups of students: those who had positive gains on pre-posttests and those who had negative gains on pre-posttests, and we examine how each group's CT-STEM practices developed as they engaged with the curricular unit.

Theory

Computational Literacy and Restructurations

As computational tools are becoming increasingly ubiquitous, computational thinking is becoming an essential skill for everyone, not just computer scientists or STEM professionals. Computer scientists have theoretically stressed the importance of algorithmic thinking for decades (Dijkstra 1974; Knuth 1985), but in the early 1980s, Papert (1980) presented an alternative empirical approach for investigating how children think with computers. More recently, Wing (2006) popularized the concept of *computational thinking* for K-12 education, claiming that computational thinking should be as fundamental as reading, writing, and arithmetic. She characterizes computational thinking as "thinking like a computer scientist" (2006, p. 36) and as "formulating a problem and expressing its solution(s) in such a way that a computer—human or machine—can effectively carry out" (2017, p. 8). Although Wing 2017 and others advocate for broadening participation in CT, many of the current definitions and examples are rooted in computer science culture and the term computational thinking is continually conflated with computer science and programming (Grover and Pea 2013; Israel et al. 2015). But if computational thinking is for everyone, then its definitions, examples, and fundamental components should not be limited to practices specific to computer scientists and be accessible to broader populations.

Computational tools have changed how science is practiced and have created new systems of knowledge that make learning concepts easier. But even before the invention of computers, scientists made representational changes that had significant benefits for learners. For example, diSessa (2001) considers how when Galileo was exploring the concept of uniform motion, he described the relationships among distance, velocity, and time in terms of lengthy, text-based theorems. With the invention of algebra, Galileo's theorems were transformed into a simpler representational form of distance equaling velocity times time: $d = v \times t$. This algebraic representational transformation modified a complex notion into a concept that students now learn in secondary school. This alternative representation is what Wilensky and Papert (2010) define as a *restructuration* of the domain: a change in the representational infrastructure of how knowledge is *externally expressed* in a domain which affects how knowledge is *internally encoded* in the mind. This is a powerful idea for the design of learning environments because just as algebra made Galileo's difficult concepts more accessible to the public hundreds of years ago, restructurations, particularly those involving computational tools, can make complex concepts more accessible to students today.

One example of a computational infrastructure that can help restructurate advanced science content is NetLogo, a programming language for agent-based modeling (Wilensky 1999). Agent-based approaches have been shown to be an effective tool for scientists to describe and explore phenomena and for learners to understand phenomena (Abrahamson and Wilensky 2007; Blikstein and Wilensky 2009; Sengupta and Wilensky 2009). Contrary to traditional mathematical models that use differential equations, agent-based models use a set of computational rules to model phenomena. For example, the Lotka-Volterra mathematical model is a time-dependent system of differential equations that represent predator-prey dynamics. These are composed of variables like the population sizes of predator and prey species and other parameters mathematically describing their interactions. Understanding the evolution of this system over time typically depends on an understanding of calculus. An agent-based model of the same phenomenon has different fundamental components, in this case, predator and prey agents, such as wolves and sheep. Such agents have characteristics that describe their current state and relatively simple rules that direct their actions and interactions. Rather than relying on equations to describe predator-prey phenomena, students can program rules governing individual agent behavior to explore complex macro-level patterns, such as extinction or overpopulation, that emerge from micro-level interactions between a large number of agents. Students draw on their intuitions about their own behavior in the world in order to determine the rules they program into their model. They can then run their model and test and refine their thinking. This approach to learning about population dynamics is beneficial for students who have not had the opportunity to learn algebra and calculus or have found those infrastructures to be too complex to master. Thus, what fundamentally makes NetLogo an example of a restructuration is that it alters how information is understood in a domain and in turn, provides a more accessible representation than traditional representations (Wilensky and Papert 2010).

Characterizing Computational Thinking and Learning in STEM

Berland and Wilensky (2015) claim that computational thinking is, in fact, not monolithic and deeply affected by the perspective of a person and the context in which the person uses a computational tool. The nature of computational thinking is influenced by the domain and context in which it exists, which varies from art to social sciences to STEM. In order to characterize the nature of computational thinking in STEM domains, Wilensky, Horn, and colleagues (Weintrop et al. 2016) outlined a taxonomy of CT-STEM practices. The researchers developed the taxonomy by conducting a literature review, examining the practices of teachers and students engaging in computational math and science activities, and consulting with teachers, researchers, and STEM professionals. The taxonomy was based on real-world examples of computational thinking as it was practiced in STEM research disciplines, as opposed to decontextualized practices or practices specific to computer science.

The taxonomy is comprised of four major strands: data practices, modeling and simulation practices, computational problem solving practices, and systems thinking practices. Each of the four major strands contains five to seven practices. For example, the data practices strand includes collecting data, creating data, manipulating data, analyzing data, and visualizing data. One practical application of this taxonomy was providing an operational definition of CT in STEM that was subsequently used to inform the design of curricula and assessments. For example, a 2-h *ecosystem stability* biology lesson was designed to engage students in CT-STEM practices and focused on the modeling and simulation strand of the taxonomy (Dabholkar et al. 2017). For this lesson, students explored population dynamics in a NetLogo simulation of an ecosystem and investigated population-level effects of parameters for individual organisms, such as reproduction rates, by exploring the simulation with various parameter values. Through their exploration, students learned about factors affecting the stability of an ecosystem and developed computational practices related to using and assessing models (Swanson et al. 2018).

Modeling and Measuring Computational Thinking

One key philosophy guiding the design of lessons that have been developed using the CT-STEM taxonomy is *constructionism* (Papert 1980; Papert and Harel 1991). A constructionist approach emphasizes creating objects that represent how a learner actively constructs and reconstructs their understanding of a domain (Kafai 1995). The act of construction allows the learner to guide their learning through the creation of personally meaningful and public artifacts. In many cases, the object that is being constructed is computational in nature (Brady et al. 2015; Sengupta et al. 2013; Sherin 2001; Wagh et al. 2017; Wilensky 2003). When constructed objects are computational, they are easily manipulated in multiple ways to represent conceptual ideas (Papert 1980). For example, in one study, students who used the RANDOM function in their computer code to generate random colors, numbers, or other chosen variables showed an understanding of how to apply stochastic functions to achieve desired results in their projects (Papert 1996). Thus, the creation of computational objects not only has the potential to represent domain knowledge but also has the affordance of representing such knowledge in multiple forms.

When learners have access to various representations of concepts, they make decisions about how to connect among these different representations and pieces of their knowledge. The more connections a learner makes between objects, the richer their understanding of the underlying concepts related to that object and ultimately, a learner develops a high quality relationship with the object and concepts (Wilensky 1991). diSessa (1993) argues that more expert knowledge systems have more reliable and productive connections between knowledge elements than novice knowledge systems. In the novice knowledge systems, elements are fragmented, loosely interconnected, and cued inconsistently. In contrast, in the expert knowledge system, elements are coherently related, strongly connected, and cued more consistently in contexts where they are productive. Learning—the progression from novice to expert—occurs through the reorganization and refinements of connections in the knowledge system. Thus, the novice knowledge system contains the foundational building blocks that are viewed as productive for the construction of expert knowledge systems. For example, foundational elements in the novice system could be based on intuition (diSessa 1993), common sense (Sherin 2006), or personal epistemologies (Hammer and Elby 2004).

Empirically, connected networks of novice and expert knowledge systems can be visualized and analyzed through network analysis tools. In general, network analyses trace the flow of information through links and nodes. In social network analysis, for example, researchers examine patterns among people's interactions, where the nodes of the network represent people and links among the nodes represent how strongly certain people are connected. To measure connections among cognitive elements, the nodes represent the knowledge and skills of one individual and the links represent the individual's associations between knowledge. These nodes are elements identified in discourse, which could be in the form of written documents, conversations, or actions. The links are analytically determined when elements co-occur in the discourse. Researchers have shown that co-occurrences of concepts in a given segment of discourse data are good indicators of cognitive connections (Arastoopour et al. 2016; Lund and Burgess 1996).

One tool for developing such discourse networks is Epistemic Network Analysis (ENA) (Shaffer et al. 2016; Shaffer et al. 2009; Shaffer and Ruis 2017). ENA measures when and how often learners make links between domain-relevant elements during their work. It accomplishes this by measuring the co-occurrences of discourse elements and representing them in weighted network models. This means that when someone repeatedly makes a link between elements over time, the weight of the link between those elements is greater. Furthermore, ENA enables researchers to compare networks both visually and through summary statistics that reflect the weighted structure of connections (Collier et al. 2016). Thus, researchers can use ENA to model discourse networks and quantitatively compare the discourse networks of individuals and groups of people in a variety of domains (Arastoopour et al. 2014; Arastoopour and Shaffer 2013; Bagley and Shaffer 2009; Hatfield 2015; Nash and Shaffer 2013). These affordances also allow researchers to make claims about assessing student knowledge development (Arastoopour et al. 2016).

Assessing CT-STEM Practices and Competencies

CT assessments have been developed in the context of block-based programming, using tools such as Scratch (Bienkowski et al. 2015; Brasiel et al. 2017; Brennan and Resnick 2012; Grover et al. 2015; Moreno-León et al. 2017; Moreno-León et al. 2015; Portelance and Bers 2015; Seiter and Foreman 2013) and Alice (Denner et al. 2014; Werner et al. 2012; Zhong et al. 2016), game design, using tools such as AgentSheets/AgentCubes (Koh et al. 2014a; Koh, Nickerson, & Basawapatna, 2014; Webb 2010), and robotics (Atmatzidou and Demetriadis 2016; Berland and Wilensky 2015; Bers et al. 2014). The CT assessments discussed in this section focus on assessments used by researchers to evaluate and measure learning.

A popular form of assessment is performance-based tests that measure CT competencies and feature the same computational tools that students use in their curricular units. For example, Brennan and Resnick (2012) developed three sets of Scratch design scenarios increasing in complexity. Within each of these sets, students chose one of two Scratch design projects that were framed as projects created by another Scratch user. After choosing a project, students were asked to explain the functionality of the project, how he or she would extend the project, and fix a bug within the code.

These assessments, such as the ones by Brennan and Resnick (2012), are deemed as authentic because they use the same tools that used in the curriculum and are representative of practices and ways of thinking within a discipline that are applicable outside of the classroom (Shaffer and Resnick 1999). However, some issues with these assessments that use authentic tools are that they are time-consuming, subjective, and sometimes inaccurate (Grover 2017). Moreover, typically no pretest is administered and, as a result, there is no baseline comparison for making claims about growth in student learning. Without a pretest, it is not clear whether students developed CT competencies as a result of participating in an intervention. Some researchers have argued that a pretest is problematic for assessing computational thinking because students require some degree of familiarity with the software in order to engage effectively with the assessment (Webb 2010; Werner et al. 2012). In other words, students need to be familiar with a tool in order to take a pretest, but if they become familiar with a tool before they take the pretest, then we forfeit a baseline-level measure.

One solution to this problem is to design pre-post assessments that use the same tools that students use in the unit, but offer a user-friendly, customized version of the tool for assessment purposes. These versions would be designed such that students without any prior experience with the tool can still productively engage with the assessment and their CT competencies can be measured (Weintrop et al. 2014). If the tool within the assessment is appropriately designed, then a pre-post assessment will not only be measuring the change in proficiency of using the computational tool but also the change in CT competencies that are elicited with the use of a particular tool within a curricular unit.

In addition to administering performance-based assessments, researchers have examined final artifacts (Bers et al. 2014; Moreno-León et al. 2015) or the use of different CT practices/competencies over time (Koh et al. 2014a; Koh et al. 2014b). Although Grover (2017) advocates the use of multiple and multi-faceted measures, most studies do not consider such measurements holistically. In one recent study that is most aligned with our work, Basu et al. (2014) designed a CT measurement approach for an ecology curricular unit using the CTSiM platform. The assessment combined pre-post scores and student work. In particular, they examined correlations among pre-post scores, quality of their computational models, and the evolution of their models over time. Similarly, we examined students' pre-post scores from performance-based assessments, but in our approach, we also examined the relationship between students' assessment scores and their responses to embedded assessment questions in the unit using discourse analytics.

In this study, we designed a curricular unit, *From Ecosystems to Speciation*, with learning objectives based on the CT-STEM taxonomy. In conjunction with the learning objectives, we developed pre-post assessments and embedded assessment prompts throughout the unit. We implemented this 10-day unit in one high school classroom with 121 students and conducted analyses on 41 students who responded to all pre-post questions. To score student pre-post responses, we developed rubrics and then separated students into positive and negative gain groups. We then examined the embedded curricular responses of one positive gain student and one negative gain student both qualitatively and as discourse networks. To examine learning at a larger scale, we quantitatively examined the curricular responses of all 41 students to determine how both positive gain and negative gain students developed or did not develop CT-STEM practices. When identifying student CT-STEM practices, we used the taxonomy as a guiding framework and thematic analysis (Braun and Clarke 2006) to identify student-constructed practices that fit under the broader taxonomy categories. This top-down, bottom-up approach allowed for the identification of emergent student-constructed CT-STEM practices but still within the categories of the taxonomy. The research questions in this study are as follows: (1) Do students demonstrate gains on a pre-post CT-STEM assessment after participating in *From Ecosystems to Speciation*? (2) How do students' CT-STEM practices change over time when participating in *From Ecosystems to Speciation* as represented by ENA discourse networks? (3) Are students' pre and post scores associated with particular CT-STEM practices as represented by ENA discourse networks?

Methods

Participants and Setting

From Ecosystems to Speciation is a 10-day biology unit focused on predator-prey dynamics, competition among species, carrying capacity, genetic drift, and natural selection and builds on previous ecology units for high school students (Hall and Wilensky 2017; Wilensky et al. 2012). Activities that took place online were split into lessons and each lesson consisted of 5–7 pages. Typically, on each page, students read a prompt with a description of a NetLogo (Wilensky 1999) model and suggestions for exploration. Then, students answered 2–5 embedded assessment questions on the same page on the basis of their exploration. The teacher, Ms. Santiago, facilitated student learning by walking around the classroom to further probe and talk through the assessment questions with students or offer technical assistance for exploring the models. She also conducted class-level discussions and demonstrations several times throughout the unit to check student understanding and explain concepts. On the first and last days of the unit, students take pre-post assessments. Figure 1 shows one page of lesson 2 in which students explored a model (using the drop-down menu and sliders to change parameters) and answered two embedded assessment questions.

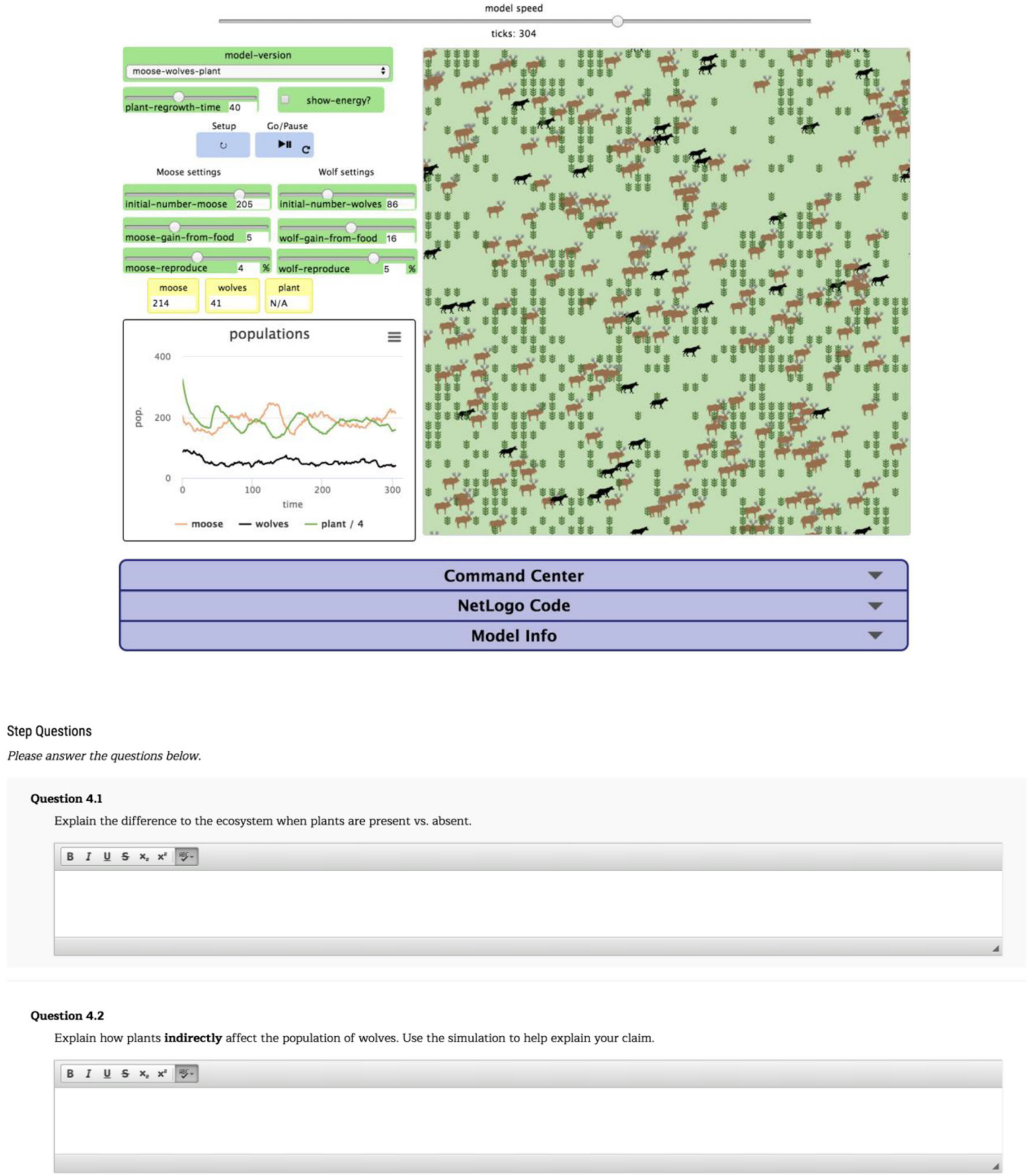
We examined students' responses to embedded assessment questions from the four lessons the students completed. The first lesson was designed for students to gather information from a real-world case study: the wolf and moose populations on Isle Royale, a uniquely isolated ecosystem in Michigan. In this lesson, students developed questions about factors that might be influencing population size changes over time and identified programmable rules to model such ecosystems. In the second lesson, students explored a NetLogo model of the Isle Royale wolf-moose ecosystem to learn about predator-prey relationships, interdependence of populations in an ecosystem, and ecosystem stability. The third lesson focused on competition between individuals in a population for resources. In this lesson, we used HubNet architecture that allows a server computer to host multiple client model (Wilensky and Stroup 1999, 2002). The teacher controlled the server model, and each student controlled an individual bug in the client models. As students engaged with the model, they learned how consumer/producer interactions for limited resources leads to a competition for those resources, even when there is no intentional effort by individuals to compete. In the fourth lesson, students moved beyond individual competition and learned how populations compete against each other by applying the concepts of stability and change in population sizes over time, direct and indirect interactions between individuals, and immediate and delayed outcomes in two different ecosystems.

Data Collection

CT-STEM units are hosted by an online platform. Students logged into their individual accounts using Chromebooks. Students' responses to online embedded assessment questions in the lessons and their pre-posttest responses were saved and anonymized.

Pre-Post Assessments

We developed two forms, A and B, for the pre-post assessments. In each form, students read a description of a NetLogo model and explored the model. Then, students answered seven questions related to the model that were aligned with CT-STEM learning objectives. The model in form A simulated the spread of contagious viruses among people (Wilensky 1998) but was redesigned to include instructions embedded in the model and the ability to change the underlying code was removed (Fig. 2). The model in form B simulated the spread of pollution (Felsen and Wilensky 2007) and the relationships among people, airborne pollution, and green landscape elements (Fig. 3).



model-version: moose-wolves-plant

plant-regrowth-time 40

□ show-energy?

initial-number-moose 205

initial-number-wolves 86

moose-gain-from-food 5

wolf-gain-from-food 16

moose-reproduce 4 %

wolf-reproduce 5 %

moose 214

wolves 41

plant N/A

Step Questions

Please answer the questions below.

Question 4.1

Explain the difference to the ecosystem when plants are present vs. absent.

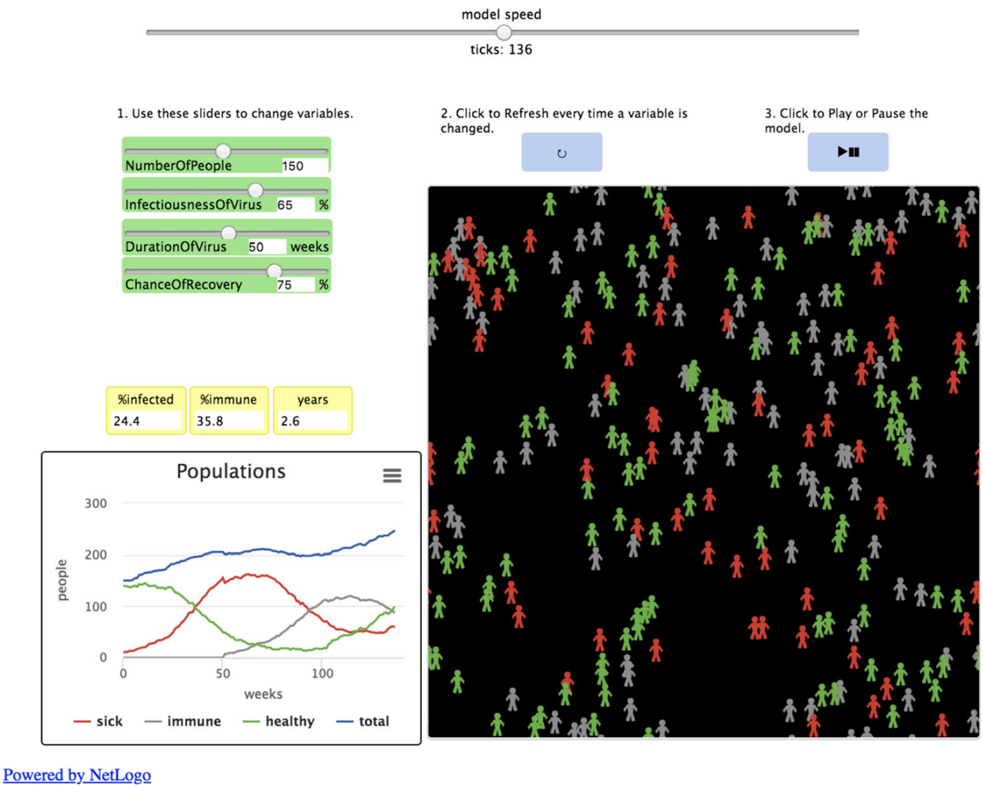
Question 4.2

Explain how plants **indirectly** affect the population of wolves. Use the simulation to help explain your claim.

Fig. 1 One page from lesson 2 in which students explored a NetLogo model of wolf-moose predator-prey relationships. In this version of the model, students added plants as agents and discovered how to stabilize the ecosystem

Both models contained three output components that were represented graphically (virus: sick, immune, and healthy people; pollution: trees, people, pollution), featured oscillations among populations of agents, and were about how people are affected by something in the environment. Students also answered almost identically worded questions on each form; the wording was only altered to identify the appropriate agents. An analysis of the pre and post responses for each question showed no significant differences between mean student scores from form A and mean student scores from form B (Table 1). For these reasons, we considered these models to be at similar difficulty levels for CT-STEM assessment purposes, and thus, form A and form B were considered to be isomorphic forms.

We randomly distributed form A to half the students for the pre assessment and form B to the remaining half. For the post assessment, the students who received form A for the pretest received form B for the posttest and those who received form B for the pretest received form A for the posttest. Because not all students completed the unit, we analyzed responses for


model speed

ticks: 136

1. Use these sliders to change variables.

2. Click to Refresh every time a variable is changed.

3. Click to Play or Pause the model.

NumberOfPeople 150

InfectiousnessOfVirus 65 %

DurationOfVirus 50 weeks

ChanceOfRecovery 75 %

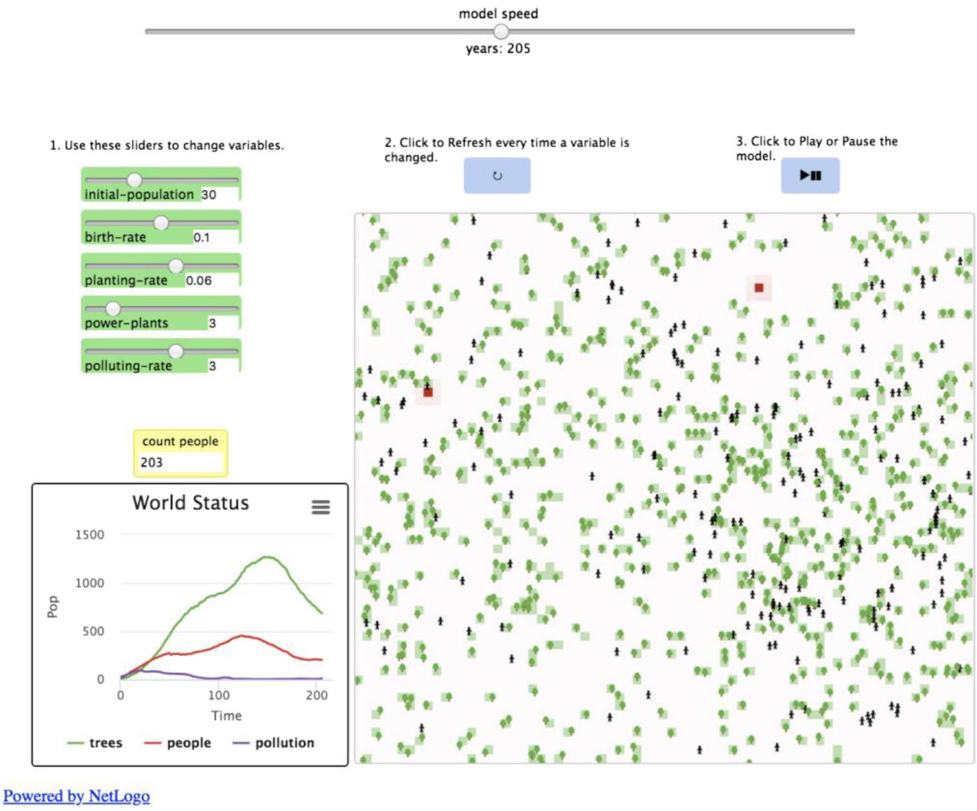
%infected 24.4

%immune 35.8

years 2.6

Fig. 2 NetLogo virus model used in form A assessment

three questions that were aligned with the main learning objectives in the lessons. Specifically, we omitted two questions asking students how changing parameters affected the model and to identify errors in code. Both questions were written at a


model speed

years: 205

1. Use these sliders to change variables.

2. Click to Refresh every time a variable is changed.

3. Click to Play or Pause the model.

initial-population 30

birth-rate 0.1

planting-rate 0.06

power-plants 3

polluting-rate 3

count people 203

Fig. 3 NetLogo pollution model used in form B assessment

Table 1 Pre and post assessment *t* test results comparing student scores on form A and form B for each question

Question	Pre or post	Form	Mean	SD	Statistic
1. Notice the oscillations (the graph moving up and down) in the graph. Why do these oscillations occur? Are there patterns in how the graph moves up and down?	Pre	A	0.97	0.71	$t(39) = 1.63$; $p > .05$
		B	0.69	0.67	
	Post	A	0.39	0.99	$t(39) = 1.32$; $p > .05$
		B	0.10	0.07	
2. List at least two ways that this model makes simplifications compared to how these viruses/pollution and other related factors behave in the real world.	Pre	A	0.37	0.50	$t(39) = 0.33$; $p > .05$
		B	0.32	0.48	
	Post	A	0.55	0.60	$t(39) = 0.18$; $p > .05$
		B	0.58	0.61	
3. Given these simplifications and your understanding of the model, why and how is this model useful for the study of viruses/pollution?	Pre	A	0.50	0.80	$t(39) = 0.56$; $p > .05$
		B	0.63	0.69	
	Post	A	0.68	0.58	$t(39) = 0.44$; $p > .05$
		B	0.77	0.69	

level that was more advanced than students had an opportunity to experience in the unit. We developed a rubric for each question based on learning objectives as well as common themes in student responses. Students received one point for every competency that was identified in the rubrics. We then summed the points across all questions for each student for their pre and post assessment. We calculated the difference scores (post minus pre) for each student. Students who decreased in their scores from pre to post were categorized as "negative gain" and those who increased in their scores from pre to post were categorized as "positive gain." Rubrics can be viewed in the Appendix.

Discourse Network Analysis of Embedded Assessment Questions in the Unit

Qualitative Coding

We used thematic analysis (Braun and Clarke 2006) to search for student responses that were related to the CT-STEM taxonomy. Braun and Clarke (2006) distinguish between a deductive top-down analysis, that is driven by theoretical frameworks and research questions, and an inductive bottom-up analysis, that is mainly data-driven and not bound to the researcher's theoretical interests.[1] Our approach used both a bottom-up analysis that allowed for identifying emergent student CT-STEM practices that were not identified a priori and also a top-down analysis in which such student practices fit broadly within the predefined taxonomy categories. In addition to reading student responses, we used word frequencies, n-grams (frequencies of phrases in the text), and topic modeling to examine the language in the data. Based on this investigation, we developed a coding scheme of seven discourse elements that were related to students' CT-STEM practices from the taxonomy (Table 2). We used this coding scheme to code student responses and the questions.

In this study, we collected all 41 students' responses to embedded assessment questions within the unit, which totaled to 1766 student responses. Because we collected such a large number of responses, we developed an automated coding algorithm to code student responses. We then used nCoder, an online software for developing and testing automated coding schemes, to test inter-rater reliability among two human coders and the automated algorithm (Eagan et al. 2017; Shaffer et al. 2015). In addition, for providing a usable platform to test inter-rater reliability, the nCoder provides a statistic, rho, that functions like a *p* value. If rho is less than .05, then the results from the sample which was coded can be generalized to a larger dataset (Shaffer 2017). To automate the coding scheme, we developed key words and regular expressions to enable automated detection for each code. For example, one regular expression for automatically coding *experimentation* includes searching for the words "to see," but not "to see *who*." We measured the reliability among two human raters and the computer. When the human and the computer disagreed, we refined the automated algorithm until we reached acceptable agreement and rho values using an unused set of student responses. Once human and the computer reached acceptable agreement kappa and rho values on a sample of data, we concluded that the code was conceptually reliable and allowed the automated algorithm to code the full dataset.

The inter-rater reliability results show that all but three pairwise agreements among rater one, rater two, and the

[1] Braun & Clark (2006) note that when researchers use a bottom-up approach, they do not completely analyze their data in an "epistemological vacuum" because they "can not free themselves [completely] of their theoretical and epistemological commitments." Even if researchers do not explicitly take a theoretical or epistemological stance, their implicit biases and points of view shape the analysis of the data.

Table 2 Coding scheme of seven CT-STEM discourse elements found in *From Ecosystems to Speciation*

CT-STEM discourse element	Definition	Student response example	Curriculum question example
Agents	Identifying agents that are used in any of the models in the unit. This does not have to be an explicit reference to the model. Examples include wolves, moose, plants, bugs, birds, and invaders.	"If there is too much wolves then there is little moose and if there is too much moose then there is little wolves."	"When a spot of green grass is eaten by your bug, what do you think you'll see happen in that spot?"
Agent actions	Describing one or more agent actions in any of the models in the unit. Examples include eating, hunting, dying, and reproducing.	"If there was another predator trying to also eat the moose there would not be as much moose for the wolves and the other predators to eat there would not be enough food for both predators."	"Were all bugs in the ecosystem equally successful at finding food? Use data to support your claim."
Biological systems	Referring to a biological phenomenon such as carrying capacity, ecosystem stability, or competition among species.	"Well from what I believe the cause to this competition is the grass because bugs needs to eat in order to gain energy but there's too many bugs so they compete each other in order to feed themselves."	"How did the outcome of this competition compare to the previous ones?"
Experimentation	Describing actions taken to experiment/explore a model. Or referring to concepts/actions related to scientific experimentation such as making and testing predictions.	"I made these changes so I could see where the two intersected quicker." "One is to make predictions and the other is to not go outside and study them one by one."	"Sketch the shape of the graph that you predict you will see for the size of the wolf population between 1959 and 2010."
Justifications	Justifying a statement or providing a reason for an action or event.	"If the moose population goes down that means the wolves are going to go down because they use moose to survive that's their food." "I changed these changes because I thought the moose would change and decrease but it did not seem to happen."	"Since moose cannot typically migrate on or off the island, what other factors might cause the size of the moose population to change from year to year?"
Quantitative amount	Using numbers to represent an amount.	"About 500 is the maximum number of moose."	NA
Temporal change	Describing a change in terms of time. May also include the description of the rate of time.	"Well for what I see the difference is that with the plants moose's population rapidly go up so fast and without the plants they still go up but after a while they start to die slowly and the wolves population go up and that makes it unstable."	"Describe the relationship between the moose and plant populations over time. Be as detailed as possible in your description."
Directional change	Describing a change and specifying the direction of change such as an increase or decrease.	"When the population stabilizes the average death rate would decrease and the average birth rate of the bugs will increase causing the population to increase even more."	"Which of the populations increase first? Explain why you think this might be the case."
Graphs	Referring to graphical forms of data from a model.	"When the graph reached its highest point the animal population did not overlap each other when one population was higher than other one was at its lowest point it goes as a cycle."	"Looking at the graph, do the peaks (highest point) of the animal populations overlap? If not describe what you see."

computer had rho values of less than .05, which means the kappa statistic from the coded sample can be generalized to the entire dataset (Table 3). Cohen's kappa values ranged from .60–1.0 and sample sizes for each code for the inter-rater reliability tests ranged from 50 to 100 excerpts.

Epistemic Network Analysis: Network Representations

After coding for CT-STEM discourse elements, we used Epistemic Network Analysis (ENA) to measure and visualize the connections students made across their discourse, as defined by the coding scheme. ENA measures the

Table 3 Inter-rater reliability (Cohen's kappa) scores for two human coders (H1 and H2) and the automated coding algorithm (computer)

Code	H1 v. H2	H1 v. computer	H2 v. computer
Agents	.92*	.92*	.84*
Agent actions	.68	.60	.84*
Biological systems	.86*	.87*	.82*
Experimentation	.83*	.94*	.88*
Justifications	.95*	.91*	.86*
Quantitative amount	.90*	1.0*	.90*
Temporal change	.89*	1.0*	.89*
Directional change	.75*	.69	.95*
Graphs	1.0*	1.0*	1.0*

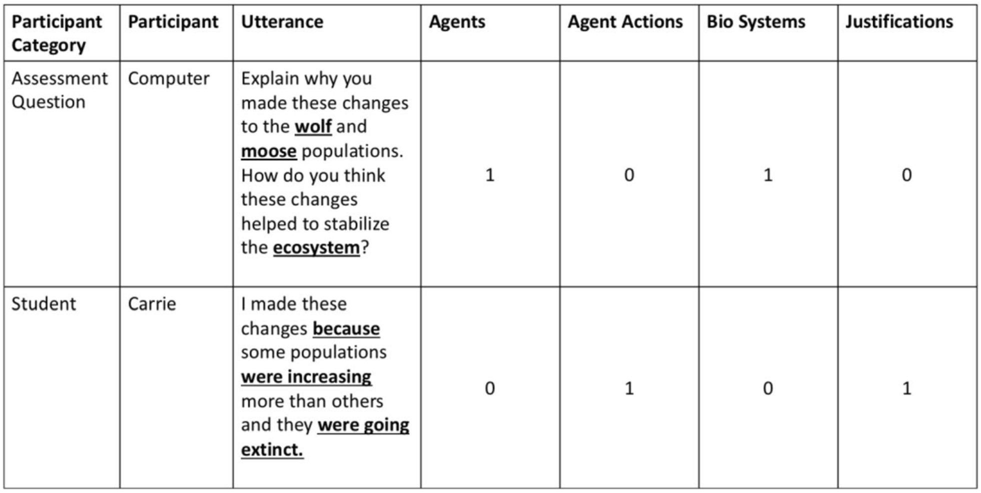
*rho < .05

connections between discourse elements, or codes, by quantifying the co-occurrence of those elements within a defined *stanza*. Stanzas are collections of utterances such that the utterances within a stanza are assumed to be closely related topically. For any two codes, the strength of their association in a network is the frequency of their co-occurrence in every accumulated stanza over time. In this study, a stanza was defined as two utterances: the embedded assessment question and the student response. Thus, co-occurrences of codes were counted if they occurred within a question, within the student's response, or between the question and the student's response. Figure 4 shows an example of one stanza for one student, Carrie. In this example, Carrie had co-occurrences within her utterance (agent actions and justifications) and also between her utterance and the assessment question (agent actions and agents, agent actions and bio systems, justifications and bio systems, and justifications and agents). We view this as a "conversation" between the curricular unit and the student.

To store the co-occurrences, ENA constructs an adjacency matrix for each stanza, which is a symmetric matrix such that both the rows and columns are codes. Every entry in the matrix represents how many times a code represented in that row co-occurs with the code represented in that column. These matrices are then summed to obtain a cumulative adjacency matrix that contains all the co-occurrences that occurred in one person's discourse over all stanzas. For example, Fig. 5 shows the cumulative adjacency matrix for Carrie.

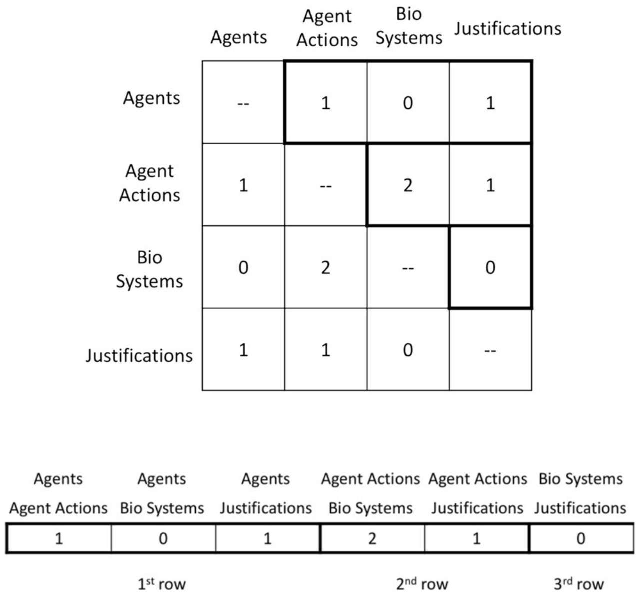
For mathematical purposes, Carrie's matrix is "unwrapped" or reshaped such that each row is appended to the one above it. Because Carrie's matrix is symmetric, only the numbers above the diagonal (the upper triangle) in the matrix are unwrapped. Carrie's unwrapped matrix is represented as a vector: [1, 0, 1, 2, 1, 0]. This vector is then converted into a normalized vector by dividing each number in the vector by its magnitude. This normalized vector would be represented as [.38, 0, .38, .76, .38, 0]. Both vectors show that the co-occurrence which occurred most frequently in Carrie's discourse was between bio systems and agent actions at a value of 2.0 and a magnitude-normalized value of .76. These values in this normalized cumulative adjacency vector are visualized as weighted links in Carrie's network (Fig. 6).

One way to interpret the weighted links is to convert the weights to percentages. In Carrie's network, the magnitude of the vector containing the normalized, weighted links can be calculated as $\sqrt{.38^2+0^2+.38^2+.76^2+.38^2+0^2}=\sqrt{1}=1$. Because the magnitude of the vector equals 1 unit, the squared components of the vector can be interpreted as percentages. For example, the squared value of the strongest weighted link in Carrie's network is $.76^2 = .58$, which means that 58% of Carrie's network is weighted towards the link between bio systems and agent actions. The remaining three connections each constitute 15% of Carrie's network.



Participant Category	Participant	Utterance	Agents	Agent Actions	Bio Systems	Justifications
Assessment Question	Computer	Explain why you made these changes to the **wolf** and **moose** populations. How do you think these changes helped to stabilize the **ecosystem**?	1	0	1	0
Student	Carrie	I made these changes **because** some populations **were increasing** more than others and they **were going extinct.**	0	1	0	1

Fig. 4 Example of one stanza in the response data of one student, Carrie



	Agents	Agent Actions	Bio Systems	Justifications
Agents	--	1	0	1
Agent Actions	1	--	2	1
Bio Systems	0	2	--	0
Justifications	1	1	0	--

Agents	Agents	Agents	Agent Actions	Agent Actions	Bio Systems
Agent Actions	Bio Systems	Justifications	Bio Systems	Justifications	Justifications
1	0	1	2	1	0
	1st row		2nd row		3rd row

Fig. 5 Carrie's cumulative adjacency matrix showing the number of co-occurrences for each pair of codes that appears in all of her discourse (top). Carrie's unwrapped cumulative adjacency matrix with only the numbers above the diagonal (bottom)

Epistemic Network Analysis: Centroid Representations

The network representations are useful when examining one, two, or three discourse networks. However, this approach is difficult when comparing many networks, and so ENA offers an alternative representation in which the centroid (center of mass) of each network is calculated and plotted in a two-dimensional space. To create a space where all networks and centroids can be equally compared, the locations of the nodes must be fixed for all networks.

In this study, the location of the nodes are determined by conducting a mean-rotation of the data in which the mean centroids of the positive gain students and the mean centroids of the negative gain students were calculated and plotted to create a line in order to maximize variance between the two groups. This line defined the first dimension (x-axis) and the mean-rotation loadings determined the location of the nodes in this first dimension (an optimization routine is also used).

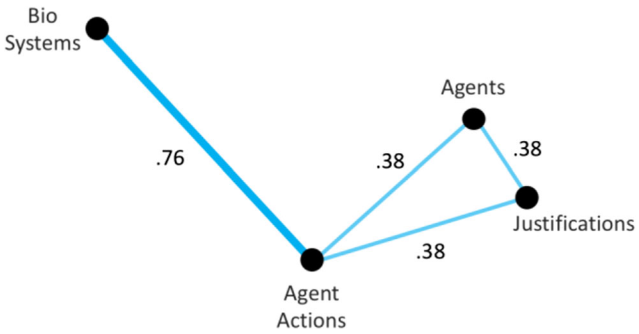


Fig. 6 Carrie's weighted discourse network representation of her cumulative adjacency matrix

The second dimension (y-axis) was calculated by performing a dimensional reduction using singular value decomposition (SVD) to rotate the vectors to show the greatest variance among the matrices and also be orthogonal to the mean-rotated first dimension. This second dimension is used for interpretation purposes so that the networks can be visualized in two dimensions. It is for interpretation purposes because the first dimension consisted of a mean-rotation in which the mean of each group is placed on the x-axis and is orthogonal to the second dimension. Because of the orthogonal restriction, there will be no differences in the means of the groups in the second dimension (for more detailed mathematical explanations of ENA, see Shaffer et al. 2016, 2009; Shaffer and Ruis 2017).

For example, Fig. 7a shows Carrie's network from above (blue) with the approximate center of mass location in a constructed two-dimensional space. Figure 7a also shows a second student's network with their approximate center of mass (red). Figure 7b shows 20 additional students' centers of mass projected into the same two-dimensional space without showing their network representations. Without examining their network representations, we can infer that the students with centers of mass that are located more to the left make more connections with bio systems and agent actions, and the students with centers of mass that are more to the right make more connections with agents and justifications. Those who have centers of mass towards the positive y-axis make more connections with bio systems and agents and those who have centers of mass towards the negative y-axis make more connections with agent actions and justifications.

Results

Pre-Post Assessments

There was a statistically significant increase from pretest (M = 1.80, SD = 1.42) to posttest (M = 2.48, SD = 1.31) scores ($t(40) = 2.38$, $p < .05$) with an effect size (Cohen's d) of .68 (Fig. 8). A Cohen's d of .68 indicates that 75% of the posttest group will be above the mean of the pretest group (Cohen's U3), 72% of the two groups will overlap, and there is a 68% chance that a person picked at random from the posttest group will have a higher score than a person picked at random from the pretest group (probability of superiority). The distribution of student pre and post score differences (post score minus pre score) ranged from − 4 to + 4 (Fig. 9).

These results indicated that (1) on average, students had learning gains from pre to post after participating in the unit and (2) the assessment was able to detect this gain.

Figure 10 shows pre and post responses for two example students. One student, Julian, had a positive gain of + 2, and one student, Pablo, had a negative gain of − 2. Both students

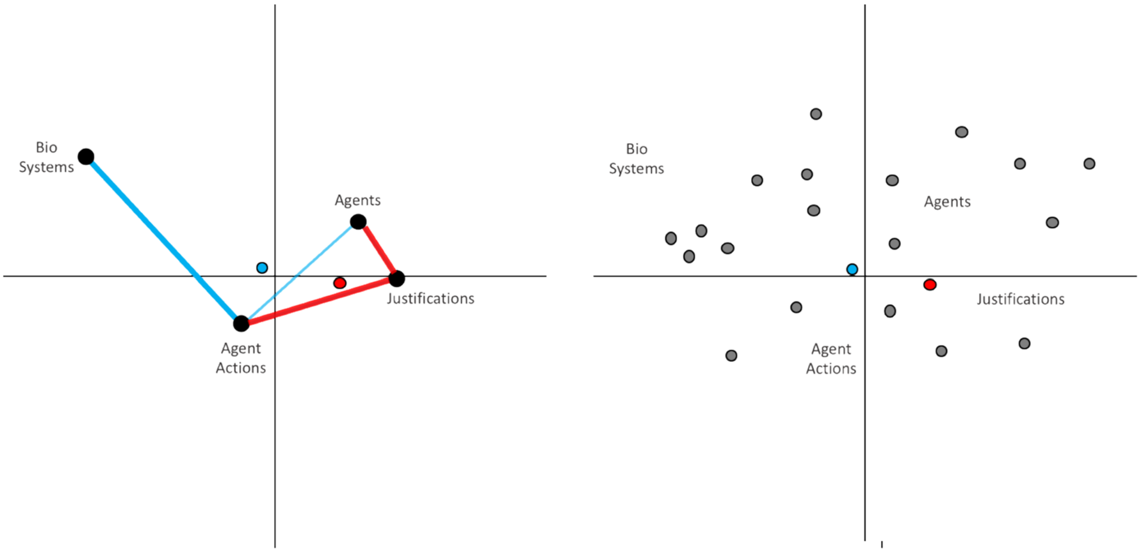


Fig. 7 **a** Carrie’s (blue) and another student’s network (red) overlaid in a two-dimensional space after a dimensional reduction on students’ normalized adjacency vectors. Approximate centers of mass are also shown for each student’s network. **b** Carrie’s (blue) and another student’s (red) approximate centers of mass along with 20 additional students (gray) in the fixed two-dimensional space that can be interpreted by the location of the nodes

received the virus model for the pretest and the pollution model for the posttest.

Curricular Activities: a Focus on Two Students

Between the pre- and posttest, students engaged with the CT-STEM biology curricular unit. Students explored models and answered questions individually on his/her own computer but were encouraged to work together. In lesson 1, students read about Isle Royale, an island in Michigan with a wolf and moose population. Students were asked to think about direct and indirect relationships among the two populations that are isolated on the island. In what follows, we focus on two students’ responses as they engaged with the curricular unit: Julian, who had positive gains from pre to post, and Pablo, who had negative gains from pre to post. Although Julian represented the majority of students who had increases from pre to post, we examined both students to get a sense of how both high and low performing students engaged in CT-STEM practices.

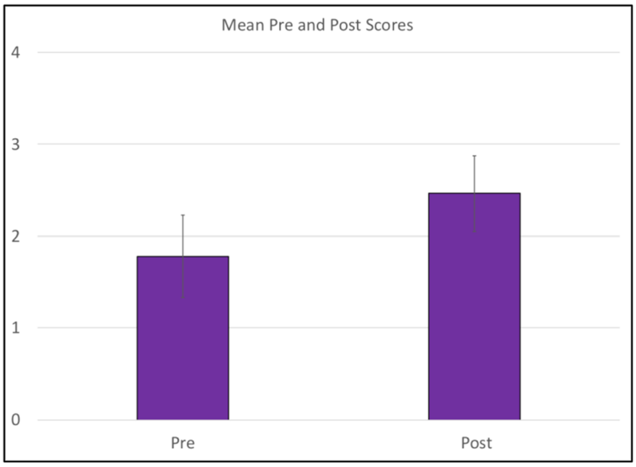


Fig. 8 Mean pre and post assessment scores for 41 students who answered all pre and post questions. Bars represent confidence intervals for a normal *t* distribution. There was a significant difference between pre and post scores ($p < .05$) with an effect size of .68

Lesson 1: Julian

Julian (positive gain) explained that the wolf population may increase when the moose populations also increases because “more wolves will be able to eat.” He also added that the wolf population may decrease later “because of the low amount of moose left on the island” indicating the effect over time of predator-prey population dynamics. Julian was able to represent his ideas in the form of oscillations on a graph (Fig. 11). Although the oscillations do not show a time lag between the two populations which is typical in predator-prey relationships, the graph shows how the size of the populations increase and decrease over time and have dependencies. Thus, Julian reasoned through the predator-prey relationships in a uniquely isolated ecosystem and provided explanations with justifications for how populations change over time.

Lesson 1: Pablo

Pablo (negative gain) also reasoned through the relationships among wolves and moose on Isle Royale, but his responses did not provide detailed information. For example, he explained that the wolf population will decrease simply “based on the limited [amount] of food there.” While his statement

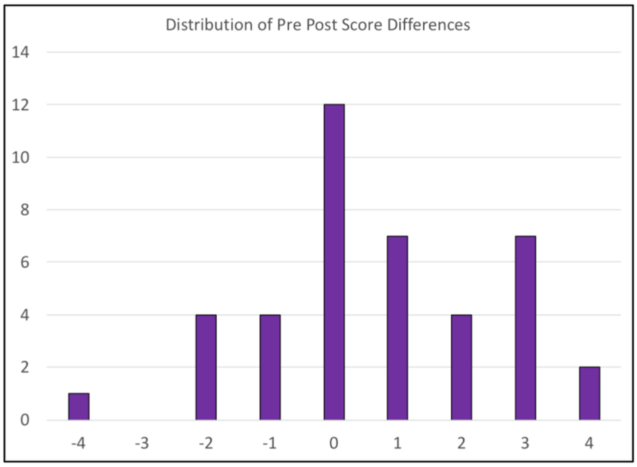


Fig. 9 Distribution of pre-post score differences (post minus pre score) for each student ranging from − 4 to + 4

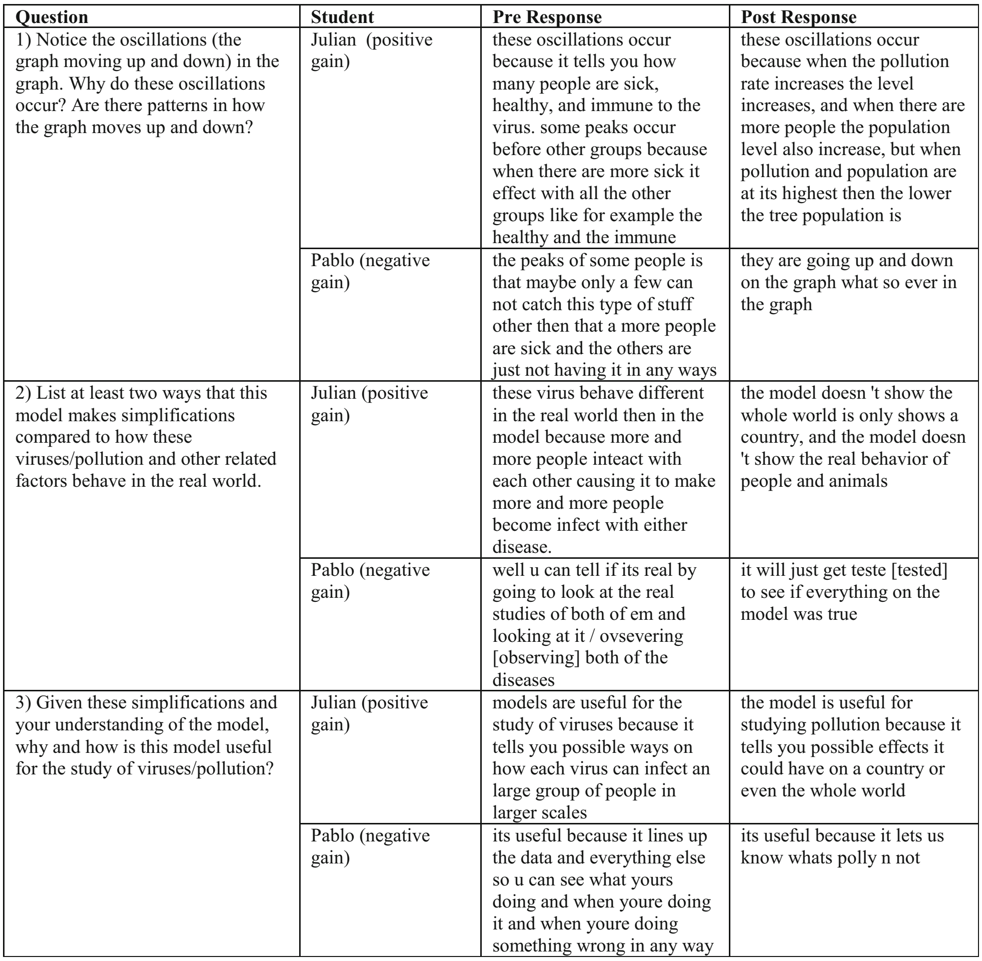
was true, Pablo did not describe the fluctuating relationships among moose and wolf populations. When asked to consider how a change in the size of population might affect another population, Pablo responded that "if that certain animal is there to [too] and just disappears or just dies period or gets eaten" then one population can affect another. Pablo's ideas were further represented in his graph in which both wolf and moose populations decreased linearly and did not fluctuate over time. Thus, as shown in his responses, Pablo identified relationships among predators and prey but did not provide descriptions about the dependencies and indirect effects among the two populations over time.

Lesson 2: Julian

In lesson 2, students examined predator-prey relationships further by using the Wolf-Moose Predation NetLogo model. This model simulates interactions between wolf and moose similar to those on Isle Royale. Using the model, students explored concepts of population stability.

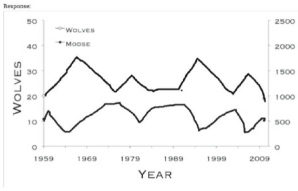
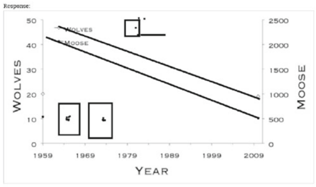
When asked about the changes he made to model and the results of his changes, Julian explained that he "increased the amount of wolves" and then explained the "moose population had at first decreased and than [then] the wolves population increased. After time pasted [past] the wolves started to quickly decrease until they died out. And because of that the moose population quickly increased." Here, Julian provided a chain of reasoning which described what he saw in the model over

Fig. 10 Julian's (positive gain student) and Pablo's (negative gain student) pre and post responses



Question	Student	Pre Response	Post Response
1) Notice the oscillations (the graph moving up and down) in the graph. Why do these oscillations occur? Are there patterns in how the graph moves up and down?	Julian (positive gain)	these oscillations occur because it tells you how many people are sick, healthy, and immune to the virus. some peaks occur before other groups because when there are more sick it effect with all the other groups like for example the healthy and the immune	these oscillations occur because when the pollution rate increases the level increases, and when there are more people the population level also increase, but when pollution and population are at its highest then the lower the tree population is
	Pablo (negative gain)	the peaks of some people is that maybe only a few can not catch this type of stuff other then that a more people are sick and the others are just not having it in any ways	they are going up and down on the graph what so ever in the graph
2) List at least two ways that this model makes simplifications compared to how these viruses/pollution and other related factors behave in the real world.	Julian (positive gain)	these virus behave different in the real world then in the model because more and more people inteact with each other causing it to make more and more people become infect with either disease.	the model doesn 't show the whole world is only shows a country, and the model doesn 't show the real behavior of people and animals
	Pablo (negative gain)	well u can tell if its real by going to look at the real studies of both of em and looking at it / ovsevering [observing] both of the diseases	it will just get teste [tested] to see if everything on the model was true
3) Given these simplifications and your understanding of the model, why and how is this model useful for the study of viruses/pollution?	Julian (positive gain)	models are useful for the study of viruses because it tells you possible ways on how each virus can infect an large group of people in larger scales	the model is useful for studying pollution because it tells you possible effects it could have on a country or even the whole world
	Pablo (negative gain)	its useful because it lines up the data and everything else so u can see what yours doing and when youre doing it and when youre doing something wrong in any way	its useful because it lets us know whats polly n not

Fig. 11 Sample of responses from Julian (positive gain) and Pablo (negative gain) in lesson 1

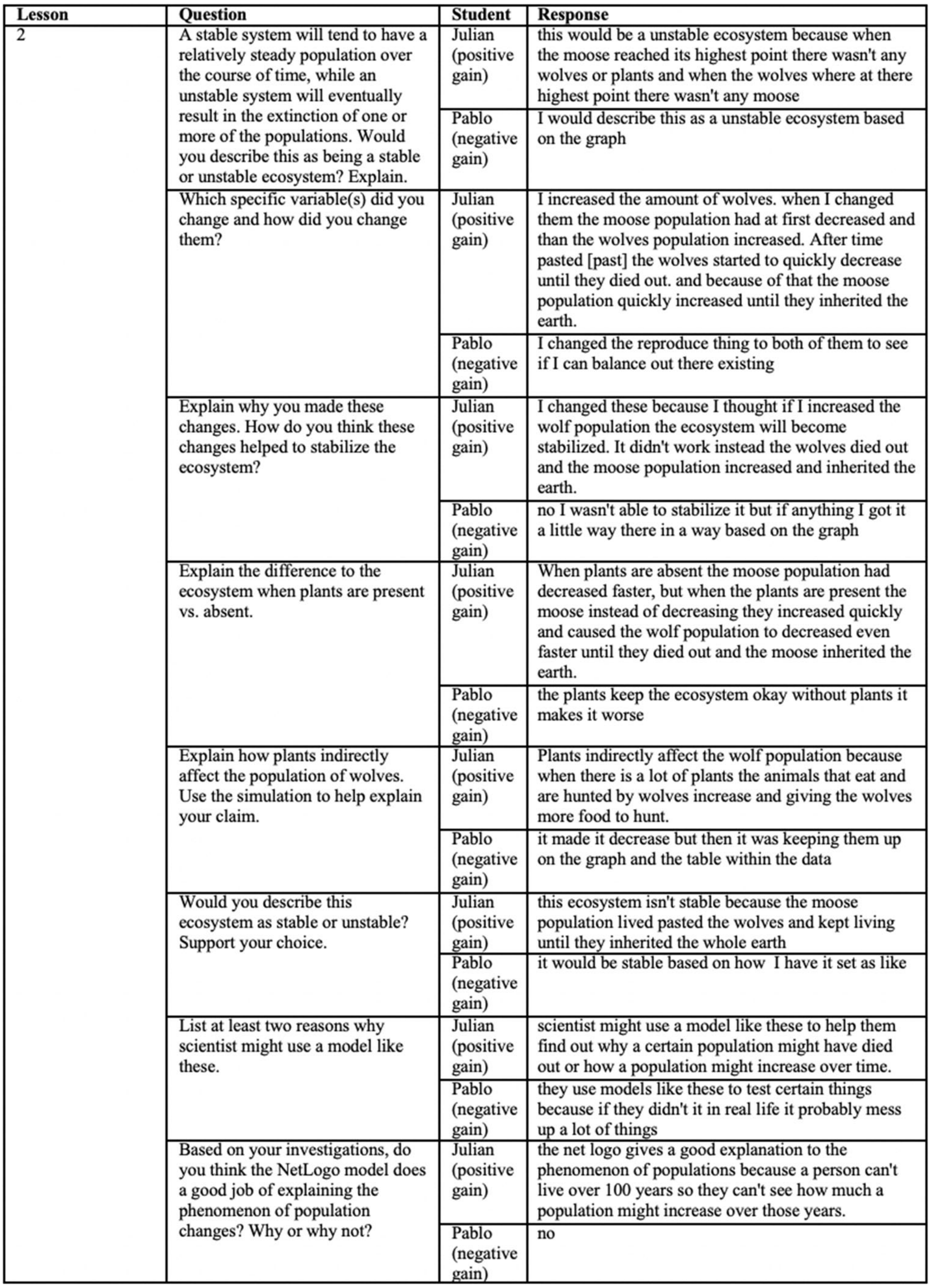
Lesson	Question	Student	Response
1	Since wolves can't typically migrate on or off the island, what other factors might cause the size of the wolf population to change from year to year?	Julian (positive gain)	the wolves population may grow because if the moose population increase the more wolves will be able to eat causing the population to increase. The wolf population may also decrease in size because of the low amount of moose left on the island
		Pablo (negative gain)	it probably decreased based on the limited [amount] of food there
	How might a change in the size of a population indirectly affect the size of another population in an ecosystem? For example, how do you think a change in the population of moose in a forest might affect the population of wolves?	Julian (positive gain)	if the size of a population is increase and the other isn't then the lower population will decrease because of the amount of wolves hunting them. But if the moose population is increased then the wolves will increase in population as well because the wolves will be able to hunt more moose and be able to feed their young.
		Pablo (negative gain)	it can affect another [population in the] ecosystem because if that certain animal is there to [too] and just disappears or just dies period or gets eaten
	Sketch the shape of the graph that you predict you will see for the size of the wolf population between 1959 and 2010. In a different color, sketch the shape of the graph that you predict you will see for the size of the moose population between 1959 and 2010.	Julian (positive gain)	
		Pablo (negative gain)	

time. When asked why he made the changes, Julian explained that he thought if he increased the wolf population then "the ecosystem will become stabilized." However, he discovered that "It didn't work instead the wolves died out and the moose population increased and inherited the earth." This shows that Julian initially predicted that increasing the wolf population would stabilize the ecosystem potentially because the wolves would eat more moose and the moose would not overpopulate. However, as Julian indicated, simply increasing the size of the wolf population did not stabilize the ecosystem.

When Julian added plants to the model, he identified an indirect relationship among plants and wolves: "when there is a lot of plants the animals that eat and are hunted by wolves increase and giving the wolves more food to hunt." Here, Julian is explaining that when there is a plentiful amount of plants, then moose will have enough food to eat and the size of the moose population will increase. As a result, the wolves will have more opportunities to hunt and eat moose. Although it is possible to adjust the parameters to make the ecosystem stable, Julian was unable to do so. However, he identified the relationships among wolves, moose, and plants and correctly described why the ecosystem was classified as unstable. At the conclusion of the lesson, Julian reflected on the use of models in scientific fields. He claimed that models are useful for scientists because "a person can't live over 100 years so they can't see how much a population might increase over those years." In other words, Julian recognized how computational models can simulate future effects and assist scientists to "find out why a certain population might have died out or how a population might increase over time."

Lesson 2: Pablo

Based on his responses at the start of the lesson, Pablo also identified the system as unstable. However, he did not provide as deep of a reasoning process as Julian. Pablo claimed, "I would describe this as a [an] unstable ecosystem based on the graph." Pablo refers to the graph as a justification for why the ecosystem is unstable but does not provide details about the size of the populations and which populations have become extinct. When asked about his changes to the model, Pablo explained that he "changed the reproduce thing to both of them" to see if he could stabilize the ecosystem but did not provide a justification for why he made changes to the reproduction parameter. He explained that he wasn't able to stabilize the ecosystem, but that he "got it a little way there in a way based on the graph." Again, Pablo refers to the graph generally to explain why the ecosystem was unstable and indicated that although the ecosystem was unstable, he was able to sustain the population longer based on the changes he made to the reproduction parameter (Fig. 12).



Lesson	Question	Student	Response
2	A stable system will tend to have a relatively steady population over the course of time, while an unstable system will eventually result in the extinction of one or more of the populations. Would you describe this as being a stable or unstable ecosystem? Explain.	Julian (positive gain)	this would be a unstable ecosystem because when the moose reached its highest point there wasn't any wolves or plants and when the wolves where at there highest point there wasn't any moose
		Pablo (negative gain)	I would describe this as a unstable ecosystem based on the graph
	Which specific variable(s) did you change and how did you change them?	Julian (positive gain)	I increased the amount of wolves. when I changed them the moose population had at first decreased and than the wolves population increased. After time pasted [past] the wolves started to quickly decrease until they died out. and because of that the moose population quickly increased until they inherited the earth.
		Pablo (negative gain)	I changed the reproduce thing to both of them to see if I can balance out there existing
	Explain why you made these changes. How do you think these changes helped to stabilize the ecosystem?	Julian (positive gain)	I changed these because I thought if I increased the wolf population the ecosystem will become stabilized. It didn't work instead the wolves died out and the moose population increased and inherited the earth.
		Pablo (negative gain)	no I wasn't able to stabilize it but if anything I got it a little way there in a way based on the graph
	Explain the difference to the ecosystem when plants are present vs. absent.	Julian (positive gain)	When plants are absent the moose population had decreased faster, but when the plants are present the moose instead of decreasing they increased quickly and caused the wolf population to decreased even faster until they died out and the moose inherited the earth.
		Pablo (negative gain)	the plants keep the ecosystem okay without plants it makes it worse
	Explain how plants indirectly affect the population of wolves. Use the simulation to help explain your claim.	Julian (positive gain)	Plants indirectly affect the wolf population because when there is a lot of plants the animals that eat and are hunted by wolves increase and giving the wolves more food to hunt.
		Pablo (negative gain)	it made it decrease but then it was keeping them up on the graph and the table within the data
	Would you describe this ecosystem as stable or unstable? Support your choice.	Julian (positive gain)	this ecosystem isn't stable because the moose population lived pasted the wolves and kept living until they inherited the whole earth
		Pablo (negative gain)	it would be stable based on how I have it set as like
	List at least two reasons why scientist might use a model like these.	Julian (positive gain)	scientist might use a model like these to help them find out why a certain population might have died out or how a population might increase over time.
		Pablo (negative gain)	they use models like these to test certain things because if they didn't it in real life it probably mess up a lot of things
	Based on your investigations, do you think the NetLogo model does a good job of explaining the phenomenon of population changes? Why or why not?	Julian (positive gain)	the net logo gives a good explanation to the phenomenon of populations because a person can't live over 100 years so they can't see how much a population might increase over those years.
		Pablo (negative gain)	no

Fig. 12 Sample of responses from Julian (positive gain) and Pablo (negative gain) in lesson 2

When Pablo added plants to the model, he claimed "the plants keep the ecosystem okay without plants it makes it worse" without explaining what it means for the ecosystem to become worse. At the conclusion of the lesson, Pablo reflected on the use of models in scientific fields. He said that scientists "use models like these to test certain things because if they didn't it in real life it probably mess up a lot of things." Overall in this lesson, Pablo identified an unstable ecosystem, described changes he made to the model to affect stability, and realized that models can be used for experimentation and simulation purposes. However, Pablo did not explain the relationships among wolves, moose, and plants, and how the phenomena of stability are affected by the predator-prey population relationships.

Lesson 3: Julian

In lesson 3, students participated in a NetLogo HubNet model in which they were all connected to a shared model managed by the teacher (Wilensky and Stroup 1999, 2002). In the first model, each student controlled a bug which wanders a world and eats grass to gain energy. In the second model, students did not control the bugs and instead observed automated bugs eat grass. Then, students compared histograms of energy distributions for each model. The goal of this lesson was to learn how variation can naturally arise in a population and to illustrate how competition can occur among individuals even without intent (Fig. 13).

When asked about the first model in which students were controlling the bugs, Julian responded that people were not able to get an equal amount of food "because some people had gotten more than someone else, this is because people saw it has a competition." He was able to represent the distribution of energy gained in the class by drawing a sketch of the general histogram. When asked about the second model in which the bugs were automated, Julian responded that there was still a competition occurring "because all the bugs raced to get the most amount of energy so they wouldn't die from low amount of energy" and explained that competition still occurs although it is not intentional "because at one point everyone is just trying to survive and live for many years before going to the after life." Although Julian did not explain the difference in variation of energy gained in the two models, he indicated that both intentional and unintentional competition occur in ecosystems because of organisms needing resources to survive.

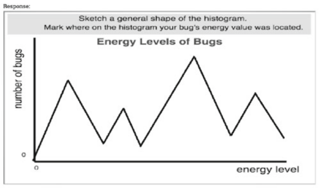
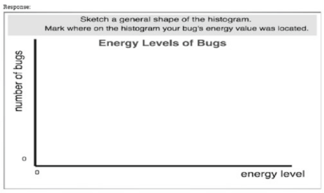
Lesson 3: Pablo

In contrast, Pablo did not specifically identify that competition nor did he explain why the student-controlled bugs were not able to receive equal amounts of food. However, he did identify that "based on how much grass there is there's really no way u [you] can get a equal amount of food." When asked to represent the distribution of energy gained in the class, Pablo left the histogram blank. At the end of the lesson, Pablo identified that competition occurred in both the student-controlled and automated bug models: "even if there [they're] not controlled we still get the same resolution." However, Pablo did not describe the difference in variation of energy gained in the two models and did not explain why competition occurred in both models.

Discourse Networks: Julian

As shown by the student responses above, Julian used data from the model to explain biological systems such as ecosystem stability and competition among individuals. As he

Fig. 13 Sample of responses from Julian (positive gain) and Pablo (negative gain) in lesson 3

Lesson	Question	Student	Response
3	Will everyone be able to get an equal amount of food in this environment? Explain your answer.	Julian (positive gain)	no because some people had gotten more than someone else, this is because people saw it has a competition.
		Pablo (negative gain)	no based on how much grass there is there's really no way u can get a equal amount of food unless every bug is on it like to save a little bit of grass
	Sketch a general shape of the histogram. Mark where on the histogram your bug's energy value was located.	Julian (positive gain)	
		Pablo (negative gain)	
	In the last exploration, bugs were not being controlled by you or anyone intentionally, but were moving about randomly. While viewing the interactions of the bugs what evidence did you notice suggesting that a competition still occurred?	Julian (positive gain)	I could tell that there was still a competition occurring because all the bugs raced to get the most amount of energy so they wouldn't die from low amount of energy.
		Pablo (negative gain)	that one was trying to get the highest number.
	Based on the model, what causes competition between individuals in an ecosystem?	Julian (positive gain)	this model does shows the competition between individual in an ecosystem because at one point everyone is just trying to survive and live for many years before going to the after life.
		Pablo (negative gain)	that everything that I controlled may not be equal or anything also n [and] even if there [they're] not controlled we still get the same resolution

progressed through the lessons, he explained relationships among the agents in the models and how models are useful for experimentation and for examining change over time.

We represented Julian's connections among computational and science concepts in his responses as discourse networks accumulated over time (Fig. 14). Julian's network from lesson 1 shows strongly weighted connections among agents, agent actions, justifications, and directional change. His lesson 1 network also showed less weighted connections between quantitative amounts, agents, and graphs. This indicates that Julian was focused on justifying agent actions in terms of their increase or decrease in population size. His lesson 2 network showed the addition of connections among bio systems, experimentation, and temporal change. This change in Julian's network occurred because in lesson 2, he made connections between agent interactions in the model and the biological concept of ecosystem stability. In lesson 3, Julian added more connections to quantitative amount and strengthened connections to experimentation and temporal change as indicated by the thicker links.

Discourse Networks: Pablo

Pablo also made connections among computational and science concepts in his responses. He focused on agents and agent actions and had few explanations or justifications for agent actions or biological phenomena. As he progressed through the lessons, he provided information as to how models are useful for simulated experimentation and that they are used to not "mess up a lot of things in real life" but did not focus on examining change over time. Pablo's connections among computational and science concepts in his responses are represented as discourse networks accumulated over time (Fig. 15). His network from lesson 1 shows higher weighted connections between agents and agent actions and additional connections among justifications, biological systems, and agents. This indicates that Pablo was focused on justifying agent actions and identifying biological phenomena. Most of the connections that Pablo made occurred in lesson 1. As he progressed through the lessons, he added minimal connections and these were mostly to experimentation because he discussed how scientists can use models for experimental purposes. His final network is more heavily weighted with agent and agent actions and less with justifications and directional/temporal change.

In addition to having different patterns of connections, both students had differences in terms of their network densities and average weighted links. Julian's network was more dense than Pablo's network. At the end of the curricular unit, Julian's discourse network had a density of .92 and Pablo's discourse

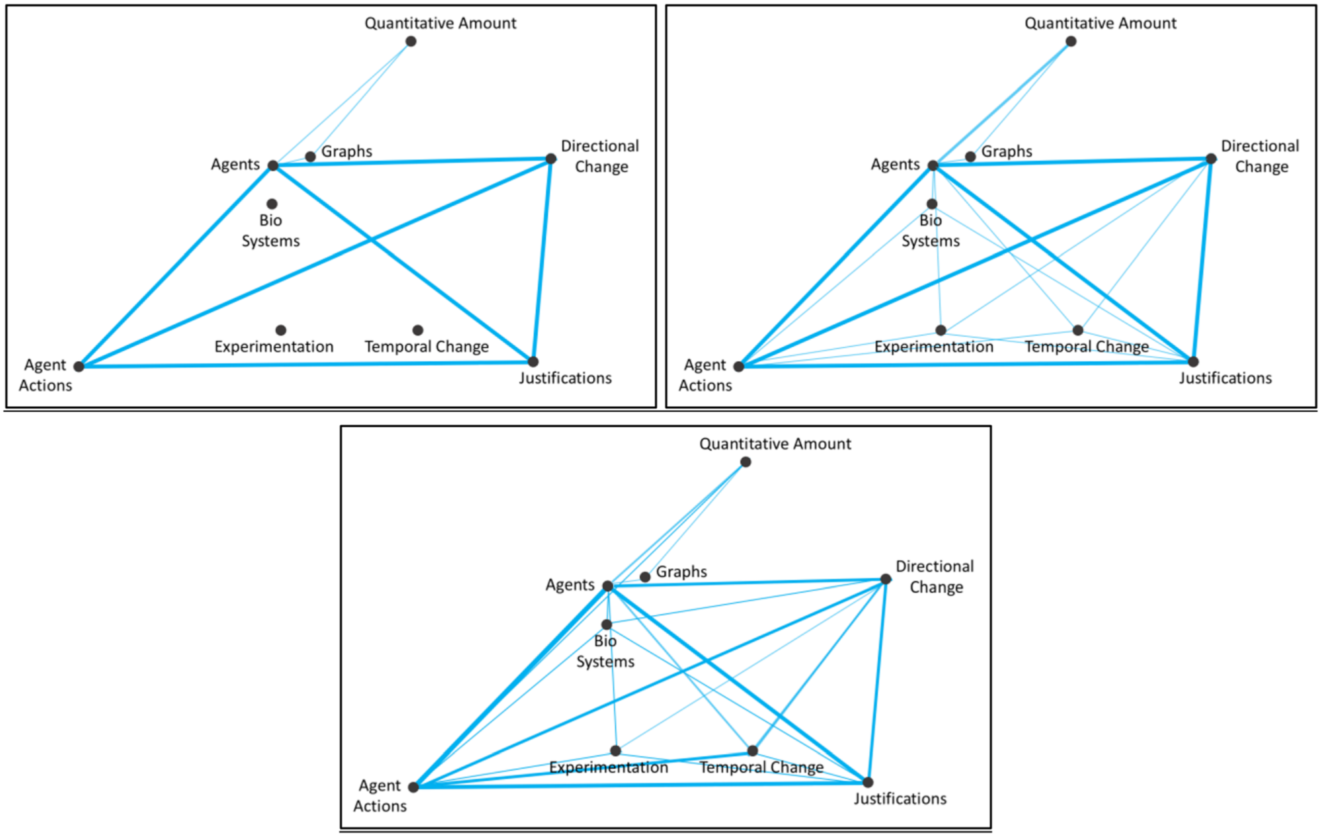


Fig. 14 Julian's (positive gain student) accumulated weighted discourse networks from lessons 1, 2, and 3

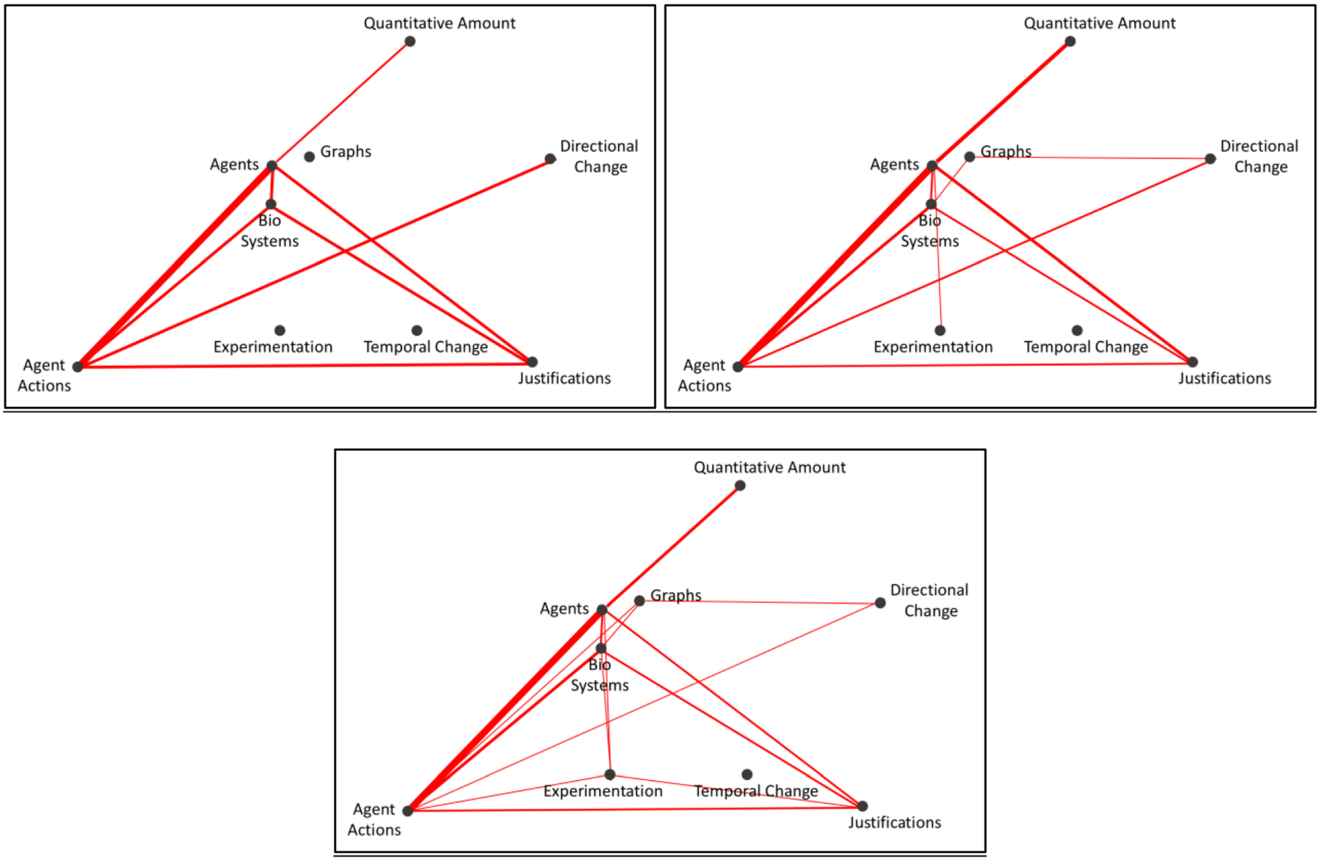


Fig. 15 Pablo's (negative gain student) accumulated weighted discourse networks from lessons 1, 2, and 3

network had a density of .54. Julian's network also had more highly weighted connections. At the end of the curricular unit, the values of the weighted links in Julian's discourse network had a mean of .12 and the links in Pablo's discourse network had a mean of .08.

Curricular Activities: All Students

In this section, we examined the discourse of all 41 students who completed the pre and post assessments. Figure 16 shows the mean discourse network for negative/zero gain students and Fig. 17 shows the mean discourse network for positive gain students. The networks show that negative/zero gain students had the strongest connections among agents, agent actions, and justifications in their networks. Positive gain students also had connections among these three elements. However, on average, positive gain students had stronger connections to justifications and also links to directional change compared to negative/zero gain students.

Figure 18 shows the subtracted mean discourse networks for positive and negative/zero gain students. The subtracted network representation shows that on average, students who had positive gains on the assessment made more connections among justifications, agents, and biological systems as well as among directional and temporal changes. In contrast, students who had negative/zero gains on the assessment made more connections with agent and agent actions and were less likely to make connections to biological systems and justifications when compared to the positive gain students.

The differences between the two networks in terms of the values of their weighted links are also shown. These values are shown for the top six largest differences between positive gain and negative/zero gain students. The largest difference between positive and negative/zero gain student networks was

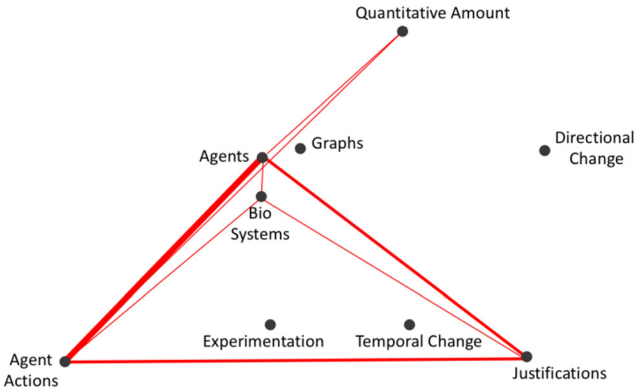


Fig. 16 Mean discourse network for negative/zero gain students. Only the strongest connections with weighted links values greater than 0.1 are shown for interpretability purposes and to show the strongest connections

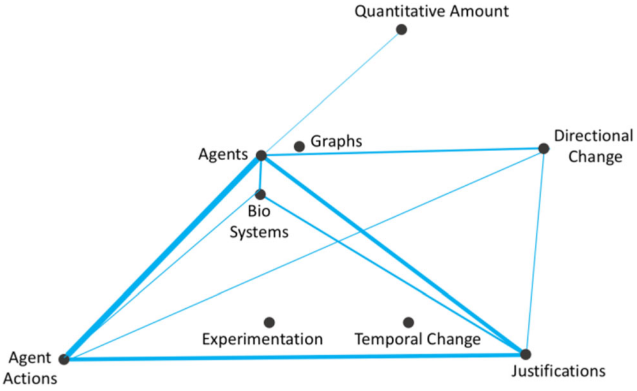


Fig. 17 Mean discourse network for positive gain students. Only the strongest connections with weighted links values greater than 0.1 are shown for interpretability purposes and to show the strongest connections

the link between agents and justifications with a value of .10 in favor of the positive gain students. The next largest difference was the link between agent actions and quantitative amount with a value of .07 and in favor the negative/zero gain students. The next four largest differences were in favor of the positive gain students between agents and directional change (.06), biological systems and justifications (.06), agents and temporal change (.05), and justifications and directional change (.05).

According to the node locations and the loading vectors, a high score on the *x*-axis represents connections to among agents, directional change, and justifications, where as a low score on the *x*-axis represents connections to agents and agent actions.

Examining the centroids of all 41 students provides a larger scale representation of the network results (Fig. 19). Positive gain students ($M = .11$, SD = .18, $n = 20$) had significantly higher discourse network centroids in the *x*-direction than negative/zero gain students ($M = -.10$, SD = .16, $n = 21$; $t(38.4) = 2.02$, $p < .05$) with an effect size of .50. Thus, positive gain students made more connections with justifications and directional/temporal changes than the negative/zero gain students.

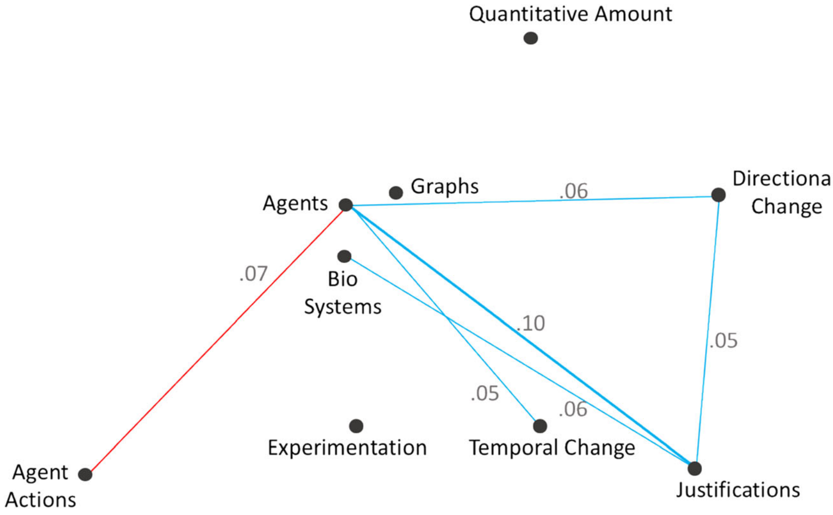
Positive gain students ($M = .45$, SD = .23) did not have significantly higher network densities than negative/zero gain students ($M = .35$, SD = .19; $t(38.12) = 1.58$, $p = .14$). However, positive gain students ($M = .07$, SD = .01) had significantly higher average weighted link values than negative/zero gain students ($M = .05$, SD = .02; $t(34.25) = 2.23$, $p < .05$) indicating that positive gain students had more strongly weighted connections among computational and science concepts and practices in their discourse networks.

Discussion

CT is an essential component of STEM education (National Research Council 2010; NGSS Lead States 2013) but has not yet been well integrated into K-12 curricula. In addition, few studies have empirically tested CT assessments or used contemporary learning sciences methods and analytics to do so (Grover and Pea 2013). In this study, we described the development of a CT-STEM biology unit for high school students and our assessment approach that used pre-post assessments to guide the analysis of the development of students' CT-STEM practices.

In this paper, we outlined the design of a pre-post assessment with two isometric forms that measured students' CT-STEM practices. There was a significant increase from pre to post which suggests that the assessment did not yield a ceiling or floor effect—the test was neither too difficult nor too easy. The significant increase also suggests that after participating in our designed CT-STEM unit, students showed gains in (1) exploring a computational model and explaining how interactions between elements produce scientific system behaviors, (2) identifying simplifications made by a computational model, and (3) assessing the match between a computational model

Fig. 18 Subtracted mean discourse networks for positive (blue) and negative/zero (red) gain students. Weighted links represent the difference between weighted links of mean positive gain student network and mean negative gain student network. The values of the weight differences are shown on links. Only the top six differences are shown for interpretability purposes and to show the highest differences



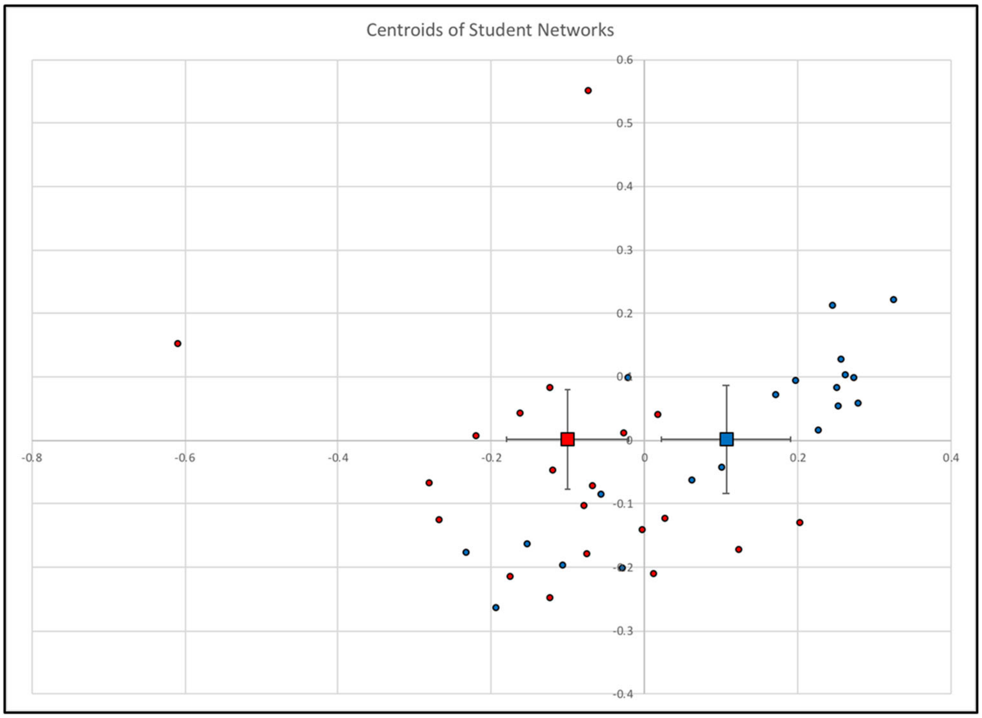


Fig. 19 Centroid of discourse networks for all 41 positive gain students (blue) and negative/zero gain students (red). Plot shows a significant difference between positive and negative/zero gain students in their discourse networks

and the phenomenon and understanding a model's range of applications. Thus, after participating in a science unit that contained computational activities, students showed learning gains on an assessment that measured CT-STEM practices. One noteworthy aspect of this pre-post assessment design is that it is one of the few performance-based assessments measuring computational thinking in a science context which includes both a pre and a post component. The pre component is critical for making claims about the effectiveness of the intervention as well as to normalize for initial level of knowledge (Adams and Wieman 2011).

However, solely examining the pre-post scores did not reveal *how* students developed CT-STEM practices when engaging with the curricular unit. To model the development of student learning, we presented an in-depth analysis of one positive gain student's and one negative gain student's responses to embedded assessments. This analysis showed different trajectories for learning in which students explored agent-based models, developed rules for agents, explained emergent biological phenomena from computational models, analyzed data from computational models, and explained affordances of computational models for science. One student who had positive gains from pre to post, Julian, supported his understanding of biological phenomena by using evidence from the model. For example, in lesson 2, when Julian explored a computational model, he was able to test his hypothesis that increasing the size of the wolf population would combat the high moose population and stabilize the ecosystem. However, after continuing to explore the model, he realized a better approach for stabilizing the ecosystem was to add plants to the model and was able to justify his approach and explain some biological mechanisms behind stability. In contrast, when Pablo explored the computational model, he varied the reproduction parameter in an attempt to stabilize the ecosystem. To determine whether the ecosystem reached stability, Pablo mainly based his reasoning on the oscillations in the graph but did not provide descriptions of the mechanisms behind stability.

To increase the scale of this result, we analyzed all 41 student responses which revealed that on average, students with positive gains were more likely to (1) provide justifications for agent actions in the model, (2) link these justifications to biological phenomena, such as ecosystem stability and competition among species, and (3) describe changes in biological computational models both temporally and directionally. In contrast, negative gain students were more likely to (1) discuss agents in terms of their actions but less likely to provide justifications and (2) link experimentation with agents, agent actions, and quantitative amounts. These results suggest that modeling connections among CT-STEM discourse elements provided a quantifiable and measurable representation of students' developing CT-STEM practices which differentiated between positive and negative gain students. These findings align with other studies that used ENA to measure differences among groups of learners (Arastoopour et al. 2014, 2016; Hatfield 2015; Knight et al. 2014; Siebert-Evenstone et al. 2017).

Our two main claims in this paper are that (1) on average, students exhibited science and computational learning gains after engaging with a science unit with computational models and (2) that the use of embedded assessments and discourse analytics tools reveals how students think with computational tools throughout the unit. The main computational tools in this unit were NetLogo agent-based models about ecosystem stability and competition among species. The results provide evidence that students used these models to make sense of biological phenomena in ways that are different from traditional equation modeling. Students connected micro-level agent actions, such as eating and reproducing, to macro-level biological system phenomena, such as ecosystem stability and extinction. In this sense, the computational activities in this unit are a restructuration of the domain—altering knowledge representations and providing opportunities for a broader range of learners to have access to scientific concepts (Wilensky and Papert 2010). Although on average, students exhibited gains, many students exhibited a decrease from pre to post. For future work, it would be helpful to more deeply analyze data from negative gain students to identify at a smaller grain size where students might be more successful with earlier intervention. For example, if students are having difficulty exploring the models, the teacher may play a more explicit role in modeling how to engage with NetLogo. Or, as the results suggest, if negative gain students are having difficulty making connections with exploring the model and analyzing temporal or directional change within the model, then more instructional supports can be employed to help students make those connections.

The results also show various student sense-making processes as they engage with the unit. This method aligns with one particular aspect of a constructionist pedagogical approach—that students may take multiple trajectories when making sense of concepts (Papert 1980). Thus, another contribution of this work is providing various models for how students develop CT-STEM practices in a science context, which may be helpful for refining existing conceptualizations including the CT-STEM taxonomy (Weintrop et al. 2016) used in this study.

One challenge in this study was the limited number of students (41 out of 121 students) who completed the pretest, curricular unit, and posttest. In the future, we will investigate the main causes of this limitation and revise our curricular unit and assessments so that a higher percentage of students can successfully complete the materials. Another limitation is that no additional data was collected about students' prior knowledge. We did not collect information about students' experiences with programming or computational thinking. Moreover, we do not have information about students' language or writing abilities, which potentially could have impacted our findings. Thus, we cannot control for such experiences. For future studies, we will collect this information, as well as collect and analyze multiple forms of data from students such as their oral participation in the classroom and their actual interactions with the models (clickstream data).

A future potential of this work is transforming discourse learning analytics into real-time assessments for instructors to use in classrooms. In this study, the analyses were completed after students completed the curricular unit. However, one avenue for future work is to conduct discourse learning analytics as students are engaging with the unit. We imagine that an assessment system for teachers could be designed that includes network analytics visualizations that are interpretable and actionable. Teachers can interpret student networks and determine how to intervene with students to assist in their development of CT-STEM practices. For example, even without such analytics Ms. Santiago, Pablo's teacher was able to see that Pablo's responses were somewhat sparse and that he may require some feedback or assistance. However, in a classroom setting, it may be difficult for a teacher to read responses, determine the content of the responses, and decide which aspects of a student's response to use as a starting point for discussion or intervention. If such discourse network analytics were available to Ms. Santiago, she could quickly view Pablo's network, see what connections Pablo was able to make with reproduction rates and oscillating populations in the wolf-moose ecosystem model, and then use what Pablo already knows about reproduction rates to begin a discussion about the mechanisms that cause oscillations in the graph and the factors that contribute to ecosystem stability beyond reproduction rates.

In another hypothetical example, without such analytics, Ms. Santiago could see that Julian's responses were longer and more detailed than most students. But again, it may be difficult to read and analyze responses in a classroom setting. Using discourse network analytics, Ms. Santiago could identify the key contributions in Julian's responses, use the assessment system to highlight such contributions, and share his response with the class as an example or as a starting point for a discussion.

In addition to on-the-fly teacher interventions, teachers could view student networks over the course of the unit to examine students' development of broader concepts. For example, without such analytics, Ms. Santiago may read through individual student responses and "code" each student response for science and computational thinking practices such as identifying agent actions or explaining biological systems. If Ms. Santiago had access to student discourse networks, she could reduce the time she spends on examining student responses and see how her students make connections across such practices over time and whether such connections change at various points in the unit. In other words, she can rely on the analytics to group her students' language together into relevant categories and provide a high-level visualization of their science learning and computational practices. Based on the visualization results, the teacher can choose which student responses to read and assess further. Looking even further into the future, if sufficient data is collected over time about student learning and teacher interventions, such assessment systems could also provide intervention suggestions and discussion prompts to engage students and thus augment a teacher's abilities to facilitate science learning for their students.

Acknowledgments This study was funded in part by the National Science Foundation (grants CNS-1138461, CNS-1441041, and DRL-1020101), the Spencer Foundation (Award #201600069), the National Science Foundation (DRL-1661036, DRL-1713110), the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals. Thank you to Daisy Rutstein from SRI for providing feedback on the development of the pre-post assessments.

Compliance with Ethical Standards

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee. Informed consent was obtained from all individual participants included in the study.

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

Appendix

Table 4 Rubrics used to score student pre and post assessments. Students received one point for each item on the rubric.

Question	Category	Description	Example
Question 1: Notice the oscillations (the graph moving up and down) in the graph. Why do these oscillations occur? Are there patterns in how the graph moves up and down?	Key Features	Describing key features of oscillations such as moving up and down or having a repeating pattern.	"The oscillations go up and down but even out almost at the end."
	Cause and Effect	Describing an event and the cause for an event. This could be an input and an output in the model.	"this happens because once everybody is better from getting sick they all become immune."
	Graph	Referring to the graph as a source of data or evidence.	"This probably occur because some groups in the graph depend on the other results to make a decision"
Question 2: List at least two ways that this model makes simplifications compared to how these viruses/pollution and other related factors behave in the real world.	Agent-based Simplification	Describing simplifications that are directly related to agents.	"people talk and touch each other"
	Non Agent-based Simplification	Describing simplifications that are not directly related to agents.	"it just shows the population"
	Missing Elements	Identifying missing elements in the model.	"The model only effects trees and humans. Not everything else on earth."
Question 3: Given these simplifications and your understanding of the model, why and how is this model useful for the study of viruses/pollution?	Illustrate	Stating a model is useful for purposes of illustrating or simulating ideas. Illustrating can address levels, potentially recognizing the visual of a n aggregate vs. agent phenomena.	"Well it shows us how every one who has what, it shows whos effected, and it shows that which person got what."
	Experimentation	Exploring the model by changing parameters and manipulating variables. Includes testing hypotheses to explore understanding of scientific concepts. Also includes stating that scientists test hypotheses and conduct experimental studies.	"it was helpful because it was somewhat realistic to real world problems and it could help out scientist study."
	Feasibility	Addressing how feasible models are. Includes discussions about the benefit of time, lower costs, scaling benefits, lower risks, and other conveniences.	"to manipulate or make adjustments to an ecosystem without affecting the real ecosystem"
	Understanding	Stating the outcome of experimentation or exploration of the model. Includes understanding relationships at a macro and micro level, understanding mechanisms of a phenomena, and gaining understanding/learning for oneself.	"to understand its danger, to see how it works, less people die"
	Application	Describing actions taken after using the model. Includes policy changes or recommendations, environmental actions, and intervening with animal populations.	"It help us calculate whether we should build more or less and how we should help our enviroment."

References

Abrahamson, D., & Wilensky, U. (2007). Learning axes and bridging tools in a technology-based design for statistics. *International Journal of Computers for Mathematical Learning, 12*(1), 23–55.

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education, 33*(9), 1289–1312. https://doi.org/10.1080/09500693.2010.512369.

Arastoopour, G., & Shaffer, D. W. (2013). Measuring social identity development in epistemic games. In *Computer-Supported Collaborative Learning Conference, CSCL* (Vol. 1).

Arastoopour, G., Chesler, N. C., & Shaffer, D. W. (2014). Epistemic persistence: A simulation-based approach to increasing participation of women in engineering. *Journal of Women and Minorities in Science and Engineering, 20*(3). https://doi.org/10.1615/JWomenMinorScienEng.2014007317.

Arastoopour, G., Shaffer, D. W., Swiecki, Z., Ruis, A. R. R., & Chesler, N. C. N. C. (2016). Teaching and assessing engineering design thinking with virtual internships and epistemic network analysis. *International Journal of Engineering Education, 32*(3), 1–10.

Atmatzidou, S., & Demetriadis, S. (2016). Advancing students' computational thinking skills through educational robotics: A study on age and gender relevant differences. *Robotics and Autonomous Systems, 75*, 661–670. https://doi.org/10.1016/j.robot.2015.10.008.

Bagley, E. A., & Shaffer, D. W. (2009). When people get in the way: Promoting civic thinking through epistemic game play. *International Journal of Gaming and Computer-Mediated Simulations, 1*(1), 36–52.

Basu, S., Kinnebrew, J. S., & Biswas, G. (2014). Assessing student performance in a computational-thinking based science learning environment. *Springer International Publishing*, 476–481.

Berland, M., & Wilensky, U. (2015). Comparing virtual and physical robotics environments for supporting complex systems and computational thinking. *Journal of Science Education and Technology, 24*(5), 628–647. https://doi.org/10.1007/s10956-015-9552-x.

Bers, M. U., Flannery, L., Kazakoff, E. R., & Sullivan, A. (2014). Computational thinking and tinkering: Exploration of an early childhood robotics curriculum. *Computers & Education, 72*, 145–157. https://doi.org/10.1016/j.compedu.2013.10.020.

Bienkowski, M., Snow, E., Rutstein, D., & Grover, S. (2015). Assessment design patterns for computational thinking practices in secondary computer science, (December), 1–46.

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. *International Journal of Computers for Mathematical Learning, 14*(2), 81–119. https://doi.org/10.1007/s10758-009-9148-8.

Brady, C., Holbert, N., Soylu, F., Novak, M., & Wilensky, U. (2015). Sandboxes for model-based inquiry. *Journal of Science Education and Technology, 24*(2–3), 265–286. https://doi.org/10.1007/s10956-014-9506-8.

Brasiel, S., Close, K., Jeong, S., Lawanto, K., Janisiewicz, P., & Martin, T. (2017). Emerging research, practice, and policy on computational thinking. In P. J. Rich & C. B. Hodges (Eds.), *Emerging research, practice, and policy on computational thinking* (pp. 327–347). https://doi.org/10.1007/978-3-319-52691-1.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. In *Annual Meeting of the American Educational Research Association. Vancouver, B.C.* https://doi.org/10.1.1.296.6602.

Collier, W., Ruis, A. R., & Shaffer, D. W. (2016). Local versus global connection making in discourse. In *International Conference of the Learning Sciences*. Singapore.

Dabholkar, S., Hall, K., Woods, P., Bain, C., & Wilensky, U. (2017). From ecosystems to speciation. Evanston, IL: Center for Connected Learning and Computer-Based Modeling, Northwestern University.

Denner, J., Werner, L., Campe, S., & Ortiz, E. (2014). Pair programming: Under what conditions is it advantageous for middle school students? *Journal of Research on Technology in Education, 46*(3), 277–296. https://doi.org/10.1080/15391523.2014.888272.

di Sessa, A. A. (2001). *Changing minds: Computers, learning, and literacy.* Cambridge, MA: Mit Press.

Dijkstra, E. W. (1974). Programming as a discipline of mathematical nature. *The American Mathematical Monthly, 81*(6), 608–612.

diSessa, A. A. (1993). Towards an epistemology of physics. *Cognition and Instruction, 10*(2), 105–225.

Eagan, B. R., Rogers, B., Serlin, R., Ruis, A. R., Arastoopour Irgens, G., & Shaffer, D. W. (2017). Can we rely on IRR? Testing the assumptions of inter-rater reliability. In *Computer Supported Collaborative Learning*. Philadelphia, PA.

Felsen, M., & Wilensky, U. (2007). NetLogo urban suite—pollution model. Evanston, IL: Center for Connected Learning and Computer-Based Modeling, Northwestern University.

Foster, I. (2006). A two-way street to science's future. *Nature, 440*(March 23), 419. https://doi.org/10.1038/440419a.

Grover, S. (2017). Assessing algorithmic and computational thinking in K-12: Lessons from a middle school classroom. In P. J. Rich, & C. B. Hodges (Eds.), Emerging research, practice, and policy on computational thinking (1st ed.). pp. 269-288. Cham, Switzerland: Springer International Publishing

Grover, S., & Pea, R. (2013). Computational thinking in K-12: A review of the state of the field. *Educational Research, 42*(1), 38–43. https://doi.org/10.3102/0013189X12463051.

Grover, S., Pea, R., & Cooper, S. (2015). Designing for deeper learning in a blended computer science course for middle school students. *Computer Science Education, 25*(2), 199–237. https://doi.org/10.1080/08993408.2015.1033142.

Hall, K., & Wilensky, U. (2017). Ecosystem stability. Evanston, IL: Center for Connected Learning and Computer-Based Modeling, Northwestern University.

Hammer, D., & Elby, A. (2004). On the form of a personal epistemology, 169–190. https://doi.org/10.4324/9780203424964.

Hatfield, D. L. (2015). The right kind of telling: An analysis of feedback and learning in a journalism epistemic game. *International Journal of Gaming and Computer-Mediated Simulations, 7*(2), 1–23.

Israel, M., Pearson, J. N., Tapia, T., Wherfel, Q. M., & Reese, G. (2015). Supporting all learners in school-wide computational thinking: A cross-case qualitative analysis. *Computers & Education, 82*, 263–279. https://doi.org/10.1016/j.compedu.2014.11.022.

Kafai, Y. (1995). *Minds in play: Video game design as a context for children's learning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Knight, S., Arastoopour, G., Shaffer, D. W., Buckingham Shum, S., & Littleton, K. (2014). Epistemic networks for epistemic commitments. In *Proceedings of the International Conference of the Learning Sciences*. Boulder, CO.

Knuth, D. E. (1985). Algorithmic thinking and mathematical thinking. *The American Mathematical Monthly, 92*(3), 170–181.

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

Koh, K. H., Nickerson, H., & Basawapatna, A. (2014b). Early validation of computational thinking pattern analysis. In *Proceedings of the*

2014 ITICSE. Uppsala, Sweden. https://doi.org/10.1145/2591708.2591724.

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. https://doi.org/10.1145/1929887.1929902.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203–208.

Moreno-León, J., Robles, G., & Román-González, M. (2015). Dr. Scratch: Automatic analysis of scratch projects to assess and foster computational thinking. *RED. Revista de Educación a Distancia, 15*(46), 1–23. https://doi.org/10.6018/red/46/10.

Moreno-León, J., Harteveld, C., Román-González, M., & Robles, G. (2017). On the automatic assessment of computational thinking skills: A comparison with human experts. *Conference on Human Factors in Computing Systems (CHI)*, 2788–2795. https://doi.org/10.1145/3027063.3053216.

Nash, P., & Shaffer, D. W. (2013). Epistemic trajectories: Mentoring in a game design practicum. *Instructional Science, 41*(4), 745–771.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11463.

National Research Council. (2010). *Report of a workshop on the scope and nature of computational thinking*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12840.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press. https://doi.org/10.17226/13165.

NGSS Lead States. (2013). *Next generation science standards: for states, by states*. Washington, DC: The National Academies Press.

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

Papert, S. (1996). An exploration in the space of mathematics educations. *International Journal of Computers for Mathematical Learning, 1*(1), 95–123.

Papert, S., & Harel, I. (1991). Situating constructionism. In *Constructionism* (pp. 1–11). Ablex Publishing Corporation. https://doi.org/10.1111/1467-9752.00269.

Portelance, D. J., & Bers, M. U. (2015). Code and tell: Assessing young children's learning of computational thinking using peer video interviews with ScratchJr. In *Proceedings of the 14th international conference on interaction design and children* (pp. 271–274). https://doi.org/10.1145/2771839.2771894.

Schanzer, E., Fisler, K., & Krishnamurthi, S. (2018). Assessing bootstrap: Algebra students on scaffolded and unscaffolded word problems. In Proceedings of the 49th ACM technical symposium on computer science education - SIGCSE'18 (pp. 8–13). Baltimore, Maryland: ACM Press. https://doi.org/10.1145/3159450.3159498.

Seiter, L., & Foreman, B. (2013). Modeling the learning progressions of computational thinking of primary grade students. In *Proceedings of the Ninth Annual International ACM Conference on International Computing Education Research* (pp. 59–66). https://doi.org/10.1145/2493394.2493403.

Sengupta, P., & Wilensky, U. (2009). Learning electricity with NIELS: Thinking with electrons and thinking in levels. *International Journal of Computers for Mathematical Learning, 14*(1), 21–50.

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. *Education and Information Technologies, 18*(2), 351–380. https://doi.org/10.1007/s10639-012-9240-x.

Shaffer, D. W. (2017). *Quantitative ethnography.* Madison, WI: Cathcart Press.

Shaffer, D. W., & Resnick, M. (1999). Thick authenticity: New media and authentic learning. *Journal of Interactive Learning Research, 10*(2), 195–215.

Shaffer, D. W., & Ruis, A. R. (2017). Epistemic network analysis: A worked example of theory-based learning analytics. In *Handbook of learning analytics and educational data mining* (p. in press).

Shaffer, D. W., Hatfield, D., Svarovsky, G., Nash, P., Nulty, A., Bagley, E. A., et al. (2009). Epistemic network analysis: A prototype for 21st century assessment of learning. *The International Journal of Learning and Media, 1*(1), 1–21.

Shaffer, D. W., Borden, F., Srinivasan, A., Saucerman, J., Arastoopour, G., Collier, W., … Frank, K. A. (2015). The nCoder: a technique for improving the utility of inter-rater reliability statistics (Epistemic Games Group Working Paper No. 2015–01).

Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics, 3*(3), 9–45.

Sherin, B. L. (2001). A comparison of programming languages and algebraic notation as expressive languages for physics. *International Journal of Computers for Mathematical Learning, 6*(1), 1–61. https://doi.org/10.1023/A:1011434026437.

Sherin, B. L. (2006). Common sense clarified: The role of intuitive knowledge in physics problem solving. *Journal of Research in Science Teaching, 43*(6), 535–555. https://doi.org/10.1002/tea.20136.

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review, 22*, 142–158. https://doi.org/10.1016/j.edurev.2017.09.003.

Siebert-Evenstone, A. L., Arastoopour Irgens, G., Collier, W., Swiecki, Z., Ruis, A. R., & Shaffer, D. W. (2017). In search of conversational grain size: Modeling semantic structure using moving stanza windows. *Journal of Learning Analytics, 4*(3), 123–139.

Swanson, H., Arastoopour Irgens, G., Bain, C., Hall, K., Woods, P., Rogge, C., et al. (2018). Characterizing computational thinking in high school science. In *International Conference of the Learning Sciences. London, UK.*

Swanson, H., Anton, G., Bain, C., Horn, M., & Wilensky, U. (2019). Introducing and Assessing Computational Thinking in the Secondary Science Classroom. In *Computational Thinking Education* (pp. 99-117). Springer, Singapore.

Wagh, A., Cook-Whitt, K., & Wilensky, U. (2017). Bridging inquiry-based science and constructionism: Exploring the alignment between students tinkering with code of computational models and goals of inquiry. *Journal of Research in Science Teaching, 54*(5), 615–641. https://doi.org/10.1002/tea.21379.

Webb, D. C. (2010). Troubleshooting assessment: An authentic problem solving activity for it education. *Procedia-Social and Behavioral Sciences, 9*, 903–907.

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 *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, 136 LNICST, 22–25*. https://doi.org/10.1007/978-3-319-08189-2_3.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology, 25*(1), 127–147. https://doi.org/10.1007/s10956-015-9581-5.

Werner, L., Denner, J., & Campe, S. (2012). The fairy performance assessment: Measuring computational thinking in middle school. In *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education - SIGCSE '12, 215–220*. https://doi.org/10.1145/2157136.2157200.

Wilensky, U. (1991). Abstract meditation on the concrete and concrete implications for mathematics education. In *Constructionism* (pp. 193–204).

Wilensky, U. (1998). *NetLogo Virus model*. Evanston, IL: Center for Connected Learning and Computer-Based Modeling, Northwestern University.

Wilensky, U. (1999). *NetLogo*. Evanston, IL: Center for Connected Learning and Computer-Based Modeling, Northwestern University.

Wilensky, U. (2003). Statistical mechanics for secondary school: The GasLab multi-agent modeling toolkit. *International Journal of Computers for Mathematical Learning, 8*(1), 1–41.

Wilensky, U., & Papert, S. (2010). Restructurations: Reformulating knowledge disciplines through new representational forms. In *Constructionism* (pp. 1–14) Paris, France.

Wilensky, U., & Stroup, W. (1999). Learning through participatory simulations: Network-based design for systems learning in classrooms. Proceedings of the 1999 Conference on Computer Support for Collaborative Learning, (1), 80.

Wilensky, U., & Stroup, W. (2002). Participatory simulations: Envisioning the networked classroom as a way to support systems learning for all. In *Presented at the Annual meeting of the American Research Education Association*, New Orleans, LA.

Wilensky, U., Novak, M., & Wagh, A. (2012). *MSIM evolution unit*. Evanston, IL: Center for Connected Learning and Computer-Based Modeling, Northwestern University.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM, 49*(3), 33. https://doi.org/10.1145/1118178.1118215.

Wing, J. M. (2017). Computational thinking's influence on research and education for all. *Italian Journal of Educational Technology, 25*(2), 7–14. https://doi.org/10.17471/2499-4324/922.

Zhong, B., Wang, Q., Chen, J., & Li, Y. (2016). An exploration of three-dimensional integrated assessment for computational thinking. *Journal of Educational Computing Research, 53*(4), 562–590. https://doi.org/10.1177/0735633115608444.

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