Int J of Sci and Math Educ (2019) 17:217–235
https://doi.org/10.1007/s10763-017-9864-9

Secondary Students' Understanding of Ecosystems: a Learning Progression Approach

Hui Jin[1] · Hyo Jeong Shin[2] · Hayat Hokayem[3] ·
Farah Qureshi[4] · Thomas Jenkins[5]

Received: 31 May 2017 / Accepted: 15 October 2017 / Published online: 30 October 2017

Abstract This study describes how we developed a learning progression (LP) for systems thinking in ecosystems and collected preliminary validity evidence for the LP. In particular, the LP focuses on how middle and high school students use discipline-specific systems thinking concepts (e.g. feedback loops and energy pyramid) to analyze and explain the interdependent relationships in ecosystems and humans' impact on those relationships. We administrated written assessments with 596 secondary students. Based on the data, we developed and validated an LP for systems thinking in ecosystems. The LP contains four levels that describe increasingly sophisticated reasoning patterns that students commonly use to explain phenomena about interdependent relationships in ecosystems. We used a Wright Map based on the Rasch model for polyotmous data to evaluate the validity of the LP. We also used the LP to compare the performance of students from different subgroups in terms of socioeconomic status (SES), gender, and school settings. The data suggests performance gaps for students with low SES and students from urban schools, but not for other traditionally under-served or under-represented groups such as female students and students from rural schools.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10763-017-9864-9) contains supplementary material, which is available to authorized users.

✉ Hui Jin
hjin@ets.org

[1] Student and Teacher Research Center, Educational Testing Service, Princeton, NJ, USA

[2] Research and Development Division, Educational Testing Service, Princeton, NJ, USA

[3] Andrews Institute of Mathematics & Science Education, Texas Christian University, Fort Worth, TX, USA

[4] National Assessment of Educational Progress, Educational Testing Service, Princeton, NJ, USA

[5] Assessment Development, Educational Testing Service, San Antonio, TX, USA

Keywords Ecosystems · Learning progression · Performance gap · Systems thinking

Introduction

The science curriculum in the USA has long been characterized as a mile wide and an inch deep—the curriculum covers a wide range of science topics but at a surface level (Schmidt, McKnight & Raizen, 1997). Currently, a reform in science education is under way. In 2012, the National Research Council (NRC) released *A Framework for K-12 Science Education* (NRC, 2012). Based on this framework, a writing team composed of members from 26 lead states developed the new national science education standards—the Next Generation Science Standards (NGSS) (NGSS Lead States, 2013). The NRC Framework and NGSS provide "a vision for education in the sciences and engineering, in which students, over multiple years of school, actively engage their understanding of the core ideas in these fields" (NRC, 2012, pp. 8–9). This vision is called three-dimensional science learning, as it emphasizes the integration of disciplinary core ideas, crosscutting concepts, and scientific and engineering practices.

The NRC Framework also calls for using learning progressions (LPs) to organize learning goals and curriculum. Learning progressions are "descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn about and investigate a topic over a broad span of time" (NRC, 2007, p. 219). LP approaches were developed through the collaboration among researchers across developmental psychology, educational measurement, and science education (Jin, Mikeska, Hokayem, & Mavronikolas, 2017, April). We emphasize two needs regarding using LPs to guide the implementation of the NRC Framework and NGSS. First, an LP is a cognitive model; it reflects a developmental perspective that focuses on how students' informal ways of thinking and reasoning develop into scientific ones. More specifically, the scientific ways of thinking and reasoning should reflect three-dimensional learning (scientific practices, disciplinary core ideas, and crosscutting concepts) conceptualized in the NRC Framework and NGSS. During the past decade, researchers have developed LPs for science content topics (Alonzo & Steedle, 2009; Duncan, Rogat & Yarden, 2009; Furtak, 2012; Gunckel, Covitt, Salinas & Anderson, 2012; Hadenfeldt, Neumann, Bernholt, Liu & Parchmann, 2016; Jin & Anderson, 2012a, 2012b; Jin & Wei, 2014; Johnson & Tymms, 2011; Mohan, Chen & Anderson, 2009; Neumann, Viering, Boone & Fischer, 2013; Plummer & Maynard, 2014) and scientific practices (Berland & McNeill, 2010; Schwarz et al., 2009). Many of these LPs address the three dimensions of science learning, but not in an explicit manner (see Jin, Johnson, Shin, & Anderson, 2017, for details). To date, only a few studies have developed LPs that explicitly address all three dimensions of science learning (e.g. Forbes, Zangori & Schwartz, 2015). There is a need to develop more LPs that explicitly integrate scientific practices, disciplinary core ideas, and crosscutting concepts because such LPs provide effective guidance for teachers to teach NGSS.

Second, LPs are useful for teachers and policy makers to help all students achieve NGSS through productive learning pathways. Teachers can use LPs to guide instruction and assessment of NGSS in their science classrooms. Policy makers can use LPs to measure and compare students' science proficiency against NGSS at school, state, and national levels. With the US classrooms becoming increasingly diverse, a critical issue regarding using LPs is whether an LP can be an equally valid representation of student understanding across different student subgroups. Moreover, can LPs be used to detect the performance gaps among different subgroups of students?

This study is intended to address the above needs and questions in a specific context—systems thinking in ecosystems. Recent years have witnessed a brisk development of the complex system science, which creates a method distinct from the traditional analytical and reductionist approach. Systems thinking, i.e., the ability to think in a system context, by its focus on interactions and relationships among the parts, and reasoning across scales, is pivotal to understanding complex systems such as financial markets, molecules, ecosystems, and hydrologic systems. Existing research has uncovered students' understanding of individual systems thinking concepts such as feedback loops (Sweeney & Sterman, 2007) and emergent processes and properties (Chi, Roscoe, Slotta, Roy & Chase, 2012; Wilensky & Resnick, 1999), and examined how students analyze and explain a complex system in terms of its structures, behaviors, and functions (Ben-Zvi Assaraf & Orion, 2005, 2010; Dauer, Momsen, Speth, Makohon-Moore & Long, 2013; Hmelo-Silver, Marathe & Liu, 2007; Hmelo-Silver & Pfeffer, 2004). Building upon this existing literature, this study developed an LP for how students use discipline-specific systems thinking concepts to analyze and explain the interdependent relationships in ecosystems and humans' impact on those relationships. This focus allows explicit connections between the LP and three-dimensional science learning. To validate the LP, we applied the quantitative analyses based on the item response theory (IRT) models, in particular, the Rasch model (Rasch, 1960). Additionally, we examined how the LP can be used to assess the performance of students across different subgroups. Our research questions are: (1) how do students use systems thinking to analyze and explain environmental phenomena about the interdependent relationships in ecosystems and humans' impact on those relationships? (2) What validity evidence supports the LP for systems thinking in ecosystems? (3) How well do students perform in using systems thinking to analyze and explain environmental phenomena? (4) How do students from different subgroups differ in their performance against the LP?

Conceptual Framework

In line with the effort to assess and promote three-dimensional learning, we define the construct in ways that cover all three dimensions of science learning. First, as we are interested in ways of thinking that students use to understand and explain phenomena in ecosystems, we chose to focus on the practice of constructing explanations. Second, we examine students' systems thinking in one content context—interdependent relationships in ecosystems and humans' impact on those relationships. This content topic is addressed in two core ideas in the NRC Framework (NRC, 2012): LS2.A and ESS3.C. LS2.A (Interdependent Relationships in Ecosystems) is a component of a core idea in life sciences; it focuses on the complex structure of ecosystems and the dynamic interdependent relations in ecosystems. ESS3.C (Human Impacts on Earth Systems) is a component of a core idea in Earth and Space Sciences; it addresses a variety of human impacts on earth systems, including human actions that affect ecosystems. Third, one crosscutting concept addressed in the NRC Framework is systems and system models (NGSS Lead States, 2013). This crosscutting concept is about both simple systems (e.g. a mechanical system) and complex systems (e.g. ecosystems). Systems thinking is a conceptual tool to analyze and interpret phenomena in complex systems. As such, systems thinking is included in the crosscutting concept of systems and system models. Therefore, we define the construct as *using discipline-specific systems thinking concepts to analyze and explain the interdependent relationships in ecosystems and humans' impact on those relationships*. It integrates one science practice (explanation), components of two core ideas (LS2.A; ESS3.C), and components of a crosscutting concept (systems and system models).

Literature on systems thinking provides important information for identifying the systems thinking concepts that are important for understanding the interdependent relationships in ecosystems. Many studies have examined student understanding of individual systems thinking perspectives, concepts, and ideas. Hmelo-Silver and her colleagues studied experts' and novices' understanding of two complex systems, a human respiratory system, and an aquarium tank system (Hmelo-Silver et al., 2007; Hmelo-Silver & Pfeffer, 2004). They found that experts use structure, behaviors, and functions to organize their knowledge of complex systems; although many novices are aware of structures, they seldom reason about behaviors and functions of complex systems. Ben-Zvi Assaraf and Orion (2005, 2010) identified characteristics of students' systems thinking in the context of earth systems. Resnick and Wilensky (*in sequentia*) studied how students and teachers used computer modeling software to analyze emergent properties of complex systems such as traffic jams, Maxwell-Boltzman distribution, and termites constructing nests (Resnick, 1996; Resnick & Wilensky, 1998; Wilensky & Resnick, 1999). Hogan (2000) investigated how students used systems thinking to reason about food-web perturbations and pollutant effects with ecosystems, and found that students seldom recognized feedback loops and indirect relations in ecosystems. Sweeney and Sterman (2000, 2007) have studied students' understanding of feedback loops, time delay, and stock and flow relationships across natural and social systems. Based on this body of literature and our prior research (Hokayem & Gotwals, 2016; Hokayem, Ma, & Jin, 2015), we identified several systems thinking concepts that are important for understanding the interdependent relationships in ecosystems and humans' impact on those relationships:

- Indirect/distant connections. In a complex system, components that seem unrelated are actually connected indirectly. This concept of indirect/distant connection can be applied to complex ecosystems. A classic example in ecological theory is trophic cascades in which top predators regulate herbivore populations, and therefore, indirectly influence the population of primary producers (mostly plants). In particular, recent studies in science suggest that humans interrupted the ecosystems through addition or removal of top predators. The addition or removal of top predators often triggers ecological phenomena involving changes in populations of predator, prey, and producer through the food chain (e.g. Ripple, Larsen, Renkin & Smith, 2001)
- Feedback loop. The interdependent relationships in complex systems can be analyzed using the concept of the feedback loop. There are two types of feedback loops. Negative feedback loops buffer changes and maintain the stability of an ecosystem. The interaction between the predator population and the prey population is an example of negative feedback. Unlike negative feedback loops, positive

feedback loops amplify changes and make an ecosystem unstable. For example, food surplus often leads to exponential growth

- Emergent properties. Emergent properties—patterns emerging from the collective interactions of all the agents (Chi et al., 2012)—are essential characteristics of complex systems. Exponential growth, carrying capacity, and energy pyramid are emergent properties of ecosystems

Ecosystems are complex systems because they have "nested" hierarchies—subsystems at a smaller scale are combined to form a system at a larger scale. The hierarchy extends from molecules and cells to individual organisms, populations, communities, and ecosystems. The subsystems and components in ecosystems interact with each other, and their interactions are explained in terms of these three systems thinking concepts. First, all components in an ecosystem, including both biotic (e.g. producers, consumers, and decomposers) and abiotic components (e.g. sunlight, air, and water), are connected. There are interactions among these components. Second, some interactions in ecosystems create positive or negative feedback loops that either amplify or buffer changes. Third, the complex relationships and interactions create emergent properties, which are unexpected behaviors of the whole ecosystem that stem from the interactions among the living and non-living things

Method

In a prior interview study (Jin et al., 2017, April), we developed five interview tasks to investigate how middle and high school students use the three systems thinking concepts to explain real-world phenomena in ecosystems. In this study, we converted some of those interview tasks into written assessment items and developed more assessment items. In this section, we describe the participants, assessments, and data analysis.

Participants

We administrated a computer-delivered assessment with 298 middle school students (grade 6 to grade 8) and 298 high school students (grade 9 to grade 12). Among these 596 participating students, 542 students provided demographic information (Table 1). In the USA, suburban schools tend to have highly qualified teachers, rigorous curricula, and high student performance. Urban schools often face challenges such as low resources, high teacher turnover, and low student performance. Rural schools tend to be small. Many rural schools are situated in remote and poor areas. They are dealing with similar challenges facing urban schools, such as poverty and diversity. In the USA, students from low-income households are eligible for free or reduced school lunches. Therefore, school lunch status is often used as an indicator of socioeconomic status (SES). African American students and Hispanic/Latino students tend to have lower performance in national and school tests (Lee & Luykx, 2007). As such, school settings, race, and SES are important demographic categories.

Table 1 Demographic data from participating students

Demographic categories	Demographic data		
	Number of students		Percentages of students (%)
School level	High schools	298	50.0
	Middle schools	298	50.0
School setting	Urban schools	43	7.9
	Suburban schools	80	14.8
	Rural schools	419	77.3
Ethnicity	American Indian or Alaskan Native	2	0.4
	Asian or Asian American	8	1.5
	Black or African American	55	10.1
	Hispanic/Latino	66	12.2
	Native Hawaiian or other Pacific Islander	2	0.4
	White	409	75.5
Gender	Male	263	48.5
	Female	279	51.5
SES	Receiving free or reduced lunch	289	53.3
	Not receiving free or reduced lunch	253	46.7

Assessment

As elaborated above, the construct is using discipline-specific systems thinking concepts to analyze and explain the interdependent relationships in ecosystems and humans' impacts on those relationships. This construct integrates the three dimensions of science learning, including the practice of constructing explanations, the crosscutting concepts of systems and system models, and two core ideas about ecosystems and human impacts on ecosystems. A set of assessment items was developed to assess this construct. All these items require students to explain real-world phenomena about relationships in ecosystems (e.g. loss of vegetation in Yellowstone National Park, reindeer population in a remote island, and hare populations and lynx populations in Canada). Scientific explanations of those phenomena require using the three key systems thinking concepts: distant/indirect connections, feedback loops, and emergent properties. As such, the items assess the integration of the three dimensions of science learning. The item pool contains 12 items, including four constructed-response items and eight two-tier items. The two-tier items require students to choose an option and then explain their choice.

The Yellowstone item is an example of the constructed-response item. It was designed based on a scientific investigation conducted in the Yellowstone National Park (Ripple et al., 2001). It is a two-tier item that assesses to what extent students identify the indirect relationships between top predators (wolves) and plants:

> By 1930, humans had killed all the wolves in Yellowstone National Park. In the 1990s, scientists found that aspen trees in the park had disappeared and vegetation along the riverbanks had vanished. One hypothesis for these changes was

> that the disappearance of the wolf population caused the plant populations to decrease. Explain how the disappearance of the wolf population might have caused the decrease of plant populations.

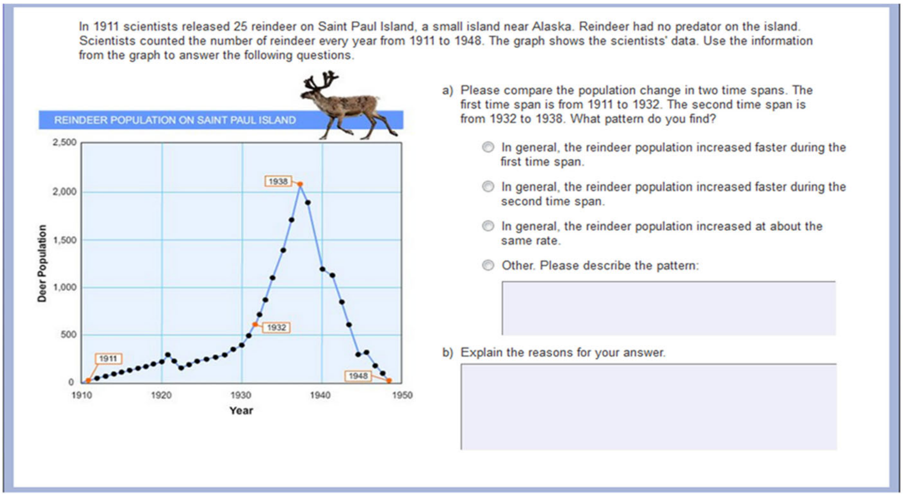
The Reindeer item is an example of the two-tier item. It consists of a pair of questions about exponential growth (Fig. 1) and a pair of questions about carrying capacity (Fig. 2). Exponential growth and carrying capacity are emergent properties of ecosystems.

We sorted the 12 items into two test forms. The high school form contains 11 items and the middle school form contains nine items. Eight common items are used to link these two test forms. These test forms were delivered using computers. Before the assessment administration, the teachers instructed students on how to use the computer delivery system. Teachers did not report problems and issues regarding students' use of the computer delivery system.

Data Analyses

As elaborated above, our research questions fall into three categories: developing and revising the LP, validating the LP, and using the LP to measure student performance. Therefore, we conducted data analyses for each category.

Developing and Revising the LP In the first step, we conducted developmental coding to develop an initial LP and associated coding rubrics. First, we selected a sample of 20 responses for each item. For each item, we sorted similar responses into groups and summarized students' main ideas for each group. Some items are two-tier items that contain a first tier that requires students to choose one option and a second tier that asks students to explain their choice. For those items, we summarized students' ideas based on students' explanation of the choice.



In 1911 scientists released 25 reindeer on Saint Paul Island, a small island near Alaska. Reindeer had no predator on the island. Scientists counted the number of reindeer every year from 1911 to 1948. The graph shows the scientists' data. Use the information from the graph to answer the following questions.

a) Please compare the population change in two time spans. The first time span is from 1911 to 1932. The second time span is from 1932 to 1938. What pattern do you find?

○ In general, the reindeer population increased faster during the first time span.

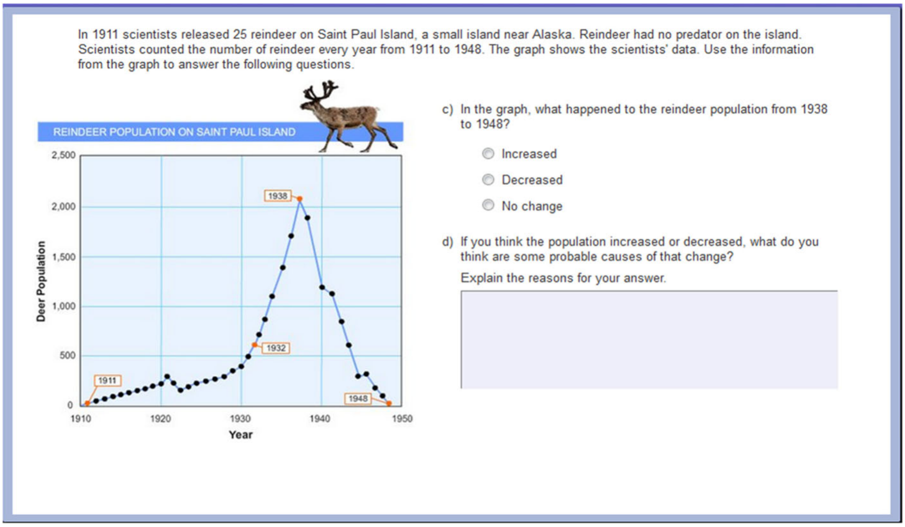
○ In general, the reindeer population increased faster during the second time span.

○ In general, the reindeer population increased at about the same rate.

○ Other. Please describe the pattern:

b) Explain the reasons for your answer.

Fig. 1 Reindeer item (screen 1)



In 1911 scientists released 25 reindeer on Saint Paul Island, a small island near Alaska. Reindeer had no predator on the island. Scientists counted the number of reindeer every year from 1911 to 1948. The graph shows the scientists' data. Use the information from the graph to answer the following questions.

c) In the graph, what happened to the reindeer population from 1938 to 1948?

○ Increased

○ Decreased

○ No change

d) If you think the population increased or decreased, what do you think are some probable causes of that change?

Explain the reasons for your answer.

Fig. 2 Reindeer item (screen 2)

Next, we compared those summaries across the groups to identify patterns. An example is provided in Table 2. For the Yellowstone item, students provided various reasons that explained how the disappearance of the wolf population might cause the decrease of plant population. We summarized the main ideas from students and sorted those ideas into three patterns.

After we identified patterns for each item, we compared those patterns across the items and collapsed and combined similar patterns. Based on this work, we generated the levels for the LP. Using this LP, we developed a more specific rubric for each item. The rubric contains detailed descriptions and examples for each level of the LP. (The

Table 2 Example of identifying patterns from students' responses

No.	Patterns	Summaries of students' ideas
1	The student does not provide any ideas about the relationships in ecosystems.	Irrelevant responses I do not know.
2	The student describes needs of individual organisms/populations or direct relationships between two organisms, but the student does not identify the distant relationships between plants and wolves.	The feces and dead bodies of the wolves were fertilizers for the plants to live and grow. The woods was a habitat for wolves. When wolves disappeared, the trees in the woods disappeared. The wolves once provided oxygen/carbon dioxide for the plants to live and grow. The wolves once spread seeds that helped the plants reproduce. The wolves and the plants must be connected because they were in one system.
3	The student describes the distant relationships between plants and wolves in terms of the food chain: plants ➔ preys such as rabbits and deer ➔ wolves.	Wolves ate herbivores. After all wolves were killed, herbivore population increased and ate up the plants.

rubrics for Yellowstone item and Reindeer item are presented in Table 3 of the Findings section.)

In the second step, we conducted full coding. More specifically, we used the scoring rubrics to score the responses from all 596 students against the LP levels. For each item, one researcher scored all responses, while a second researcher scored 20% of the responses. When the inter-rater agreement was lower than 85% for an item, the researchers discussed and resolved the discrepancies through revising the rubric. Then the revised rubric was used to re-score the responses. For each item, we used Cohen's kappa to calculate the inter-rater reliability in the final round of scoring. The results show that Cohen's kappa ranged from 0.51 to 0.91 with the median kappa of 0.70 (Cohen, 1960), and weighted kappa ranged from 0.60 to 0.94 with the median kappa of 0.79 (Cohen, 1968). These indicate moderate to high reliability in scores for each item from different raters (Landis & Koch, 1977). Finally, we discussed the discrepancy in the final round of scoring and reached agreement.

Validating the LP In the second step, quantitative methods were used to obtain validity evidence of the LP. According to Kane (2006), validation should generate

Table 3 The learning progression for systems thinking in ecosystems

Levels	Level description	Exemplar responses for items
1. No idea	I do not know; the student does not describe any relationships among organisms.	*Yellowstone item* (Response 1): The color is missing in the 1930s.
2. Individual organisms	The student describes relationships in terms of needs of individual organisms or random causes	*Yellowstone item* (Response 2): the wolf might have been fertilaizing [fertilizing] the area around the trees and now the trees have nothing making the land fertal [fertile].

3. Relation-ships and patterns	The student identifies distant relations and patterns of interactions in ecosystems, and may attempt to use systems thinking concepts to explain a phenomenon. However, the student cannot successfully use system concepts to construct explanations.	*Yellowstone item* (Response 3):
		The disappearance of wolves follows the increase of mice, rabbits, etc., which are herbivores. This stands to the reason that these herbivores decimated the plant population first by the grass, and then with the larger trees. This happens because the grass provides nutrients for the soil, which the trees must have to survive. (The student recognizes distant relationship between top predators and plants.)
		Reindeer item (Response 4):
		(c) Decreased. (d) The reason for the decrease of the population of reindeers is because of more predators that are eating them.

Level 4: Mecha-nisms	Use systems thinking concepts to construct a causal mechanism that explains phenomena about interactions in ecosystems.	*Reindeer task*: (Response 5) (c) Decreased. (d) The animal population decreased because they probably exceeded the food supply of the island. With a population so dense, the vegetation was not able to grow fast enough to support the 2000+ reindeer on the island.

(1) an interpretive argument that "…specifies the proposed interpretation and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performance" (Kane, 2006, p. 23); and (2) a validity argument that "…provides an evaluation of the interpretive argument" (Kane, 2006, p.23). In the present study, the LP is the interpretive argument—it presents our interpretation of the reasoning patterns that students use to explain interdependent relationships in ecosystems. To support this LP, we obtain the following validity evidence. An LP is assumed to describe a sequence of qualitatively different reasoning patterns of students (i.e., the LP levels). This interpretation is generalized from scores produced based on the administrated assessments. To evaluate this assumption, we performed the IRT analyses to examine whether the LP levels are differentiated from each other (Wilson, 2009). More specifically, we fitted the partial credit model (Masters, 1982), which is the Rasch model (Rasch, 1960) for the polytomous data. In doing so, we used the test analysis modules (TAM) package provided in R software (Robitzsch, Kiefer & Wu, 2017) to estimate item difficulties and students' proficiencies on the same logit scale. We predicted weighted likelihood estimates (WLEs) as students' proficiencies. Wright Maps (Figs. 5, 6, 7, and 8) were generated to present these results. The Wright Map provides graphical evidence of the validity of the LP—whether the levels of the LP are differentiated from each other and whether the higher levels of the LP are more difficult for students to achieve.

Using the LP to Measure Student Performance In the final step, we used the LP to measure student performance. In particular, we compared students' performance using the WLEs resulting from the second step. In order to test the statistical significance of performance by variables of our interest, we conducted *t* test for two samples (e.g. male vs. female, students receiving free/reduced lunch vs. students not receiving free or reduced lunch) or analysis of variance (ANOVA) for more than two samples (e.g. urban vs. suburban vs. rural). In order to compensate multiple testing problems, we applied the Bonferroni correction method for adjusting *p* values. Given the standard significance level of 0.05 and three variables of interest, we used the conservative *p* values of 0.02. Moreover, we examined whether students' performances are statistically different for different subgroups at the item level based on the differential item functioning (DIF) analysis (Holland & Wainer, 1993). To achieve item fairness, analyses of DIF can be informative to identify problematic items (e.g. Camilli & Shepard, 1994; Sudweeks & Tolman, 1993). In this study, we conducted DIF analyses to identify possibly biased items and to examine the performance gaps by gender and socioeconomic status (SES) at the item level. We conducted DIF analysis by gender and SES but not by other variables (e.g. school settings) because the sample sizes of the gender/SES subgroups are sufficient and balanced. The analysis conducted in this step allowed us to measure student performance and to compare the performance across different subgroups.

Findings

In this section, we report the following results: the final LP, the validity evidence of the LP, and student performance.

The Final LP

The final LP contains four levels (Table 3). At level 1, students do not provide any explanation for the phenomenon. They provided "I do not know" type of responses or responses that are irrelevant to the questions. Response 1 is an example of irrelevant response.

At level 2, students reason at the level of individual organisms. They recognize the feeding relationships among directly related populations but they cannot identify any indirect relationships. For example, one item was designed based on a scientific investigation conducted in the Yellowstone National Park (Ripple et al., 2001). The item provides a brief description of an event that happened in Yellowstone National Park: "By 1930, humans had killed all the wolves in Yellowstone National Park. In the 1990s, scientists found that aspen trees in the park had disappeared and vegetation along the riverbanks had vanished. One hypothesis for these changes was that the disappearance of the wolf population caused the plants to decrease." Students are asked to explain how the disappearance of the wolf population might have caused the decrease of plant population. We found several common explanations. One common explanation is that killing wolves caused the disappearance of vegetation because wolves provided fertilizer for the trees (see Response 2). Another common explanation is that wolves and trees are in one ecosystem, and therefore, they are interdependent and the disappearance of one of them caused the disappearance of the other. However, such explanations do not specify how wolves and plants are connected.

At level 3, students are able to identify distant/indirect relationships among populations and patterns of interactions among different species. At this level, students recognize the connections between the top predators and plants. For example, the correct answer to the Yellowstone item is at Level 3—students are able to explain that killing wolves caused an increase in herbivores that destroyed plants (see Response 3). However, students relying on level 3 reasoning are not able to use complicated systems thinking concepts—feedback loops and concepts related to emergent properties such as exponential growth and energy pyramid—to explain patterns and changes in ecosystems. Take Response 4 as an example, the student recognized that reindeer population is affected by predation, but the student did not provide a satisfactory explanation of how predation caused the decrease of reindeer population. The student explained that the decrease of reindeer population was because more predators appeared. The student did not recognize that more predation happened because the large reindeer population around the year 1938 makes it easier for predators to hunt for reindeer.

At level 4, students are able to use sophisticated discipline-specific systems thinking concepts such as exponential growth, carrying capacity, energy pyramid, and feedback loop to explain phenomena. Response 4 indicates that the student recognized that the increase of reindeer population caused more competition for resources such as food supplies.

Validity Evidence of the LP

The Wright Map (Fig. 3) provides us quantified locations of item difficulty and students' performance on the same scale. This allows us to visually examine whether the levels of the systems thinking were differentiated from each other. Undifferentiated levels in the Wright Map would indicate that the scoring rubric or developed learning progression framework is not empirically supported to interpret students' understanding of systems thinking. The left side of the Wright Map displays the distribution of students' performance estimates (WLEs) while the right side represents the distribution of the Thurstonian thresholds for each item. Each item has two or three threshold values, 1, 2, or 3, representing the transition between level 1 and level 2, level 2 and level 3, and level 3 and level 4, respectively. For example, the location of the second threshold (labeled as 2) for "Item 9" in the Wright Map is close to zero logit. This represents that students with average ability (which is zero logit) have about 50% chance of transition from level 2 to level 3 for this item.

In general, the levels of the LP are differentiated from each other and clustered for the same level, indicating that the assessment items differentiated students in their understanding and assessed the same construct. This provides not only the validity evidence to support the internal structure of the assessment (Wilson, 2009) but also the validity evidence that supports the generalization assumption—students' systems thinking can be depicted in terms of an LP that contains four distinguishable levels.

Students' Overall Performance

The overall student performance across items is presented in the pie chart, using the scores of students' responses across all items (Fig. 4). For example, 455 responses were scored at level 1, which account for 7% of all responses. In general, most students used the reasoning patterns at level 2 and level 3 to reason about interdependent relationships in ecosystems. Only 3% responses achieved level 4, indicating that using emergent property concepts to explain interdependent relationships in ecosystems is challenging for most students.

Performance Gaps

Regarding performance gaps, we calculated the differences in average performance for particular subgroups, using two analyses: overall performance differences via *t* tests and ANOVA and item-level performance differences via DIF analysis. Next, we examined the item fairness for subgroups that demonstrated performance gaps.

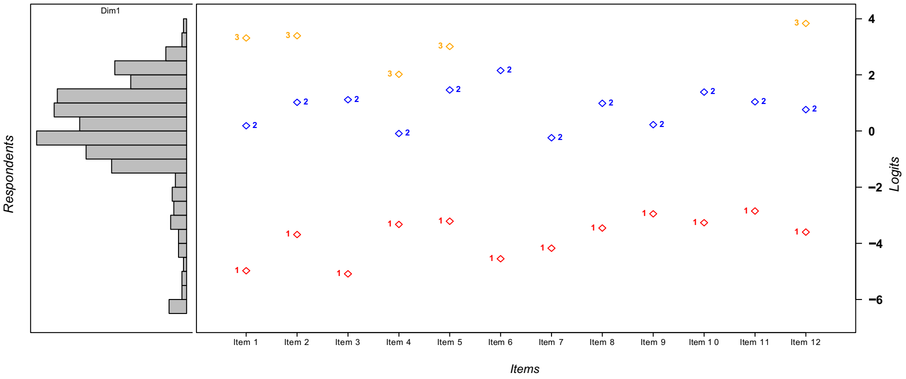


Fig. 3 Wright Map generated based on analysis of all responses

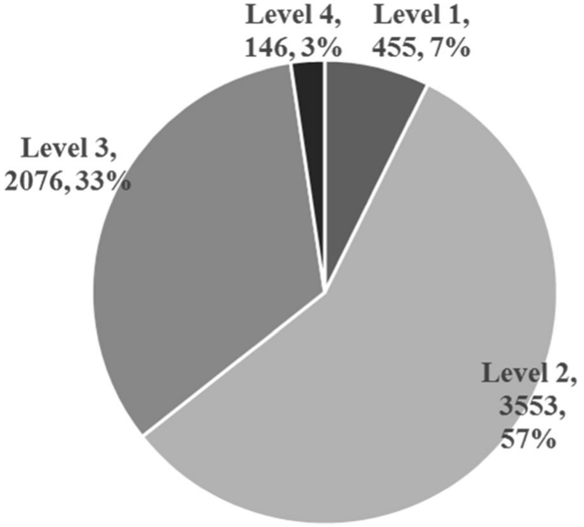


Fig. 4 Distribution of students' responses along the LP

Identification of Performance Gaps at the Overall Performance Level We compared the means of students' performance using WLEs for different subgroups using *t* tests and ANOVA. The results suggest that there is no significant difference between female students and male students (mean (female) = 0.17 (sd = 1.45) vs. mean (male) = 0.02 (sd = 1.89), $t = -1.06$, $p = 0.29$). Students who receive free/reduced lunch showed significantly lower performance than those who did not (mean (receiving free/reduced lunch) = − 0.27 (sd = 1.70) vs. mean (not receiving free/reduced lunch) = 0.42 (sd = 1.58), $t = 4.82$, $p < 0.001$). Lastly, students showed statistically different performance by school settings (mean (urban) = − 0.59 (sd = 1.15) vs. mean (suburban) = 0.08 (sd = 1.35) vs. mean (rural) = 0.17 (sd = 1.76), F = 4.14 (2539), $p = 0.02$). In particular, the average performance for students in urban areas was statistically significantly lower compared to students in rural areas ($p = 0.01$, using the Bonferroni procedure). These results suggest performance gaps for low SES students and urban school students, but not for female students and students from rural schools at the overall performance level.

Identification of Performance Gaps at the Individual Item Level Ideally, all items in the assessment are expected to measure the ability regardless of students' gender, social and cultural background, and region. However, an item may favor a particular subgroup of students because the item context is familiar to those students. Analyses of DIF can detect whether an item favors one subgroup over the other subgroups by comparing students who are at the same level of performance but come from different subgroups. For example, we can investigate if rural students have higher performance on an item than urban students even though the overall performance levels of those rural students and urban students are the same. From a technical perspective, an item is deemed to exhibit DIF, if the response probabilities for that item cannot be fully explained by the ability of the student and a set of difficulty parameters for that item.

We explored the existence of DIF with respect to gender and SES because the sample sizes of the subgroups within these two categories are sufficient (see Table 1).

Table 4 Model fit comparisons across three models

Model	Gender	SES
Model 1: partial credit model (no DIF)	9174.459	9174.459
Model 2: DIF (invariant step structure)	8292.517	8276.399
Model 3: DIF (variant step structure)	8363.141	8375.510

To examine the existence of DIF, we conducted two different analyses. One analysis assumes that the step structure (i.e., the transition between levels as indicated as thresholds) is the same for two groups, while the other assumes that the step structure is different for two groups. Table 4 presents the comparison of the model fit of the three different analyses: (1) model 1—partial credit model without DIF; (2) model 2—DIF with invariant step structure between subgroups; (3) model 3—DIF with variant step structure between subgroups.

We used the Bayes information criterion (BIC, Schwarz, 1978) to compare the model fits. Lower BIC values indicate better model fits, and the model with the lowest BIC best conforms to the data. With regard to both gender and SES, the DIF model with invariant step structure (Model 2) fitted the data best. This suggests that two items exhibited DIF (comparison between Model 1 and Model 2) but different subgroups show similar patterns in the transition between levels in the LP for those DIF items (comparison between Model 2 and Model 3).

We then interpreted the results using the best-best fitting model (Model 2). For the DIF by gender, the interaction between the item and gender provides us whether certain items were relatively easier for a certain gender, given that both genders have the same latent ability. Item 8, Item 3, and Item 9 appear to be easier for female students, while Item 6 and Item 11 appear to be easier for male students, at the 0.05 significance level. For the DIF by SES, Item 2 and Item 6 appear to be more difficult for students who received free/reduced lunch, and Item 1 and Item 11 appear to be more difficult for students who did not receive free/reduced lunch. We investigated whether these flagged

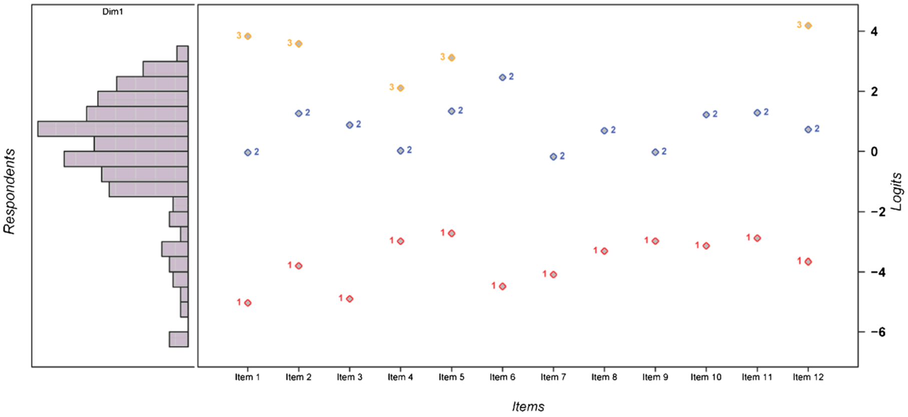


Fig. 5 Wright Maps by gender (male students)

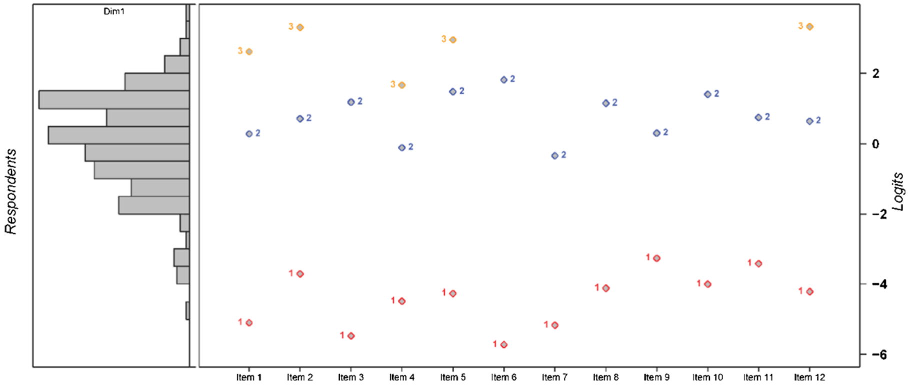


Fig. 6 Wright Maps by gender (female students)

items might show potential bias toward a certain subgroup, but we could not find any potential item content issues associated with certain subgroups.

While the above analyses show the existence of DIF in some items, it is the *magnitude of the DIF* that determines if the effect of that DIF is of substantive importance. In this sense, we found two items that are significantly more difficult than other items. First, Item 2 is significantly more difficult for students receiving free/reduced lunch, but the difference estimate is 0.26. If all of the items exhibited DIF of this magnitude, it would shift the free/reduced lunch receiving students' ability distribution by about 16% of a student standard deviation. With just one item having this magnitude of DIF, the effect is much smaller. Second, Item 6 exhibits much larger DIF than other items. In fact, if all of the items in the test had behaved like this item, the estimated mean score of students who received free/reduced lunch would be 0.62 logits lower additionally than that of students who did not receive free/reduced lunch. That is about 38% of a student standard deviation. As we examine the content of Item 2 (exponential growth of a sheep population) and Item 6 (the predator-prey relationship between fox and rabbits), we could not find information to explain the above findings of these two items. Further investigation is needed to examine student understanding of these two items.

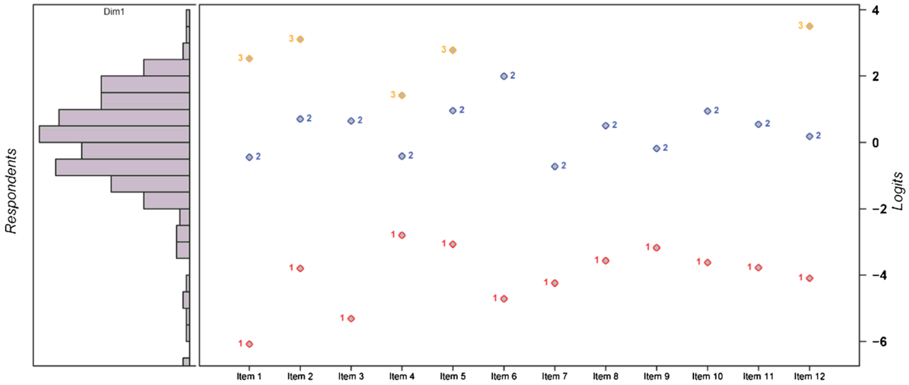


Fig. 7 Wright Maps by SES (high SES students—students who did not receive free/reduced lunch)

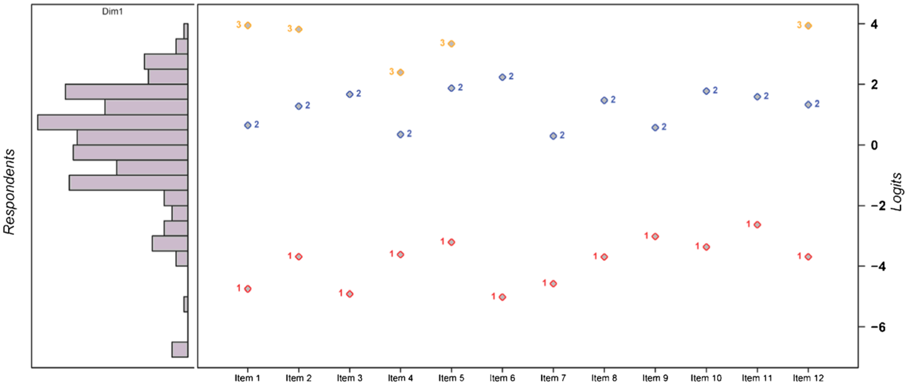


Fig. 8 Wright Maps by SES (low SES students—students who received free/reduced lunch)

In summary, we examined performance gaps associated with SES and gender at the overall test level via *t* tests and ANOVA, and at the individual item level via DIF analysis. For SES, the performance gap exists not only in the overall performance but also at the individual item level. For gender, the performance gap only exists at the individual item level. Given the magnitude of the DIF for several items, it is necessary to investigate and revise item contents in order to improve the fairness of the systems thinking assessment. To illustrate the differences between the two subgroups in gender/ SES, we fitted the PCM separately for each subgroup and generated Wright Maps. Wright Maps for gender subgroups are presented in Figs. 5 and 6. These figures show that the thresholds for male students are located higher than those for female students. Wright Maps for SES subgroups are presented in Figs. 7 and 8. These figures show that thresholds for students with low SES students are located higher than those for students with high SES. Additionally, in the Wright Maps for gender subgroups (Figs. 5 and 6), DIF items and non-DIF items show different distributions of thresholds. The distances between thresholds for DIF items (Items 3, 6, 8, 9, and 11) are slightly different between female students and male students, but the distances between thresholds for non-DIF items are almost identical for male and female students. Further qualitative research is needed to find the reason.

Conclusion and Discussion

Researchers have developed hypothetical LPs based on literature (e.g. Smith, Wiser, Anderson & Krajcik, 2006), used student assessment data to develop and validate LPs (e.g. Neumann et al., 2013), developed LPs in contexts where innovative instructional materials were used (e.g. Plummer & Krajcik, 2010), used LPs to develop effective curriculum materials (e.g. Jin, Zhan, & Anderson, 2013), developed teacher knowledge measures associated with student LPs (e.g. Jin, Shin, Johnson, Kim & Anderson, 2015), and examined how teachers used LPs in classrooms (Furtak, 2012; Furtak & Heredia, 2014; Jin et al., 2017). This study has two major contributions to the LP research.

First, most LPs were developed before the release of the NGSS and NRC framework and therefore do not explicitly address the integration of the three dimensions of science learning. Two recent studies developed learning progressions for three-dimensional science learning; these learning progressions use several variables to describe student development in "fused performance"—performances that fuse practices, crosscutting concepts, and core ideas (Forbes et al., 2015; Gotwals & Songer, 2013). This study contributes to this research effort by developing an LP for systems thinking in ecosystems. NGSS crosscutting concepts and core ideas are very broad. Therefore, a strategy used in this study is to select components of crosscutting concepts and core ideas that can be organized and integrated into a coherent construct. The assessment focuses on students' practice of constructing explanations about interdependent relationships in ecosystems and human impacts on those relationships. This focus not only connects components in two disciplinary core ideas (LS2.A Interdependent relationships in ecosystems and ESS3.C human impacts on earth systems), but also integrates these core ideas with a scientific practice—constructing explanations. Moreover, systems thinking, a component of the crosscutting concept of systems and system models, serves as the conceptual tool to understand the core ideas and construct the explanations.

Second, using LPs to detect differences in average performance among subgroups is important for evaluating and comparing learning outcomes, but it has not been sufficiently researched. In this study, we identified the same reasoning patterns from students in different subgroups, suggesting the possibility that one LP can be used with different student subgroups. We used quantitative methods to compare the performance of subgroups and found performance gaps for low SES students and urban school students, not for other traditionally under-representative groups such as female students in the sample. This result suggests that more efforts should be directed to promoting the science understanding of low SES students and urban school students.

Attention should be given to the limitations of this study. First, due to insufficient sampling, we were not able to compare performance gaps for different ethnicity groups and performance gaps for ELL (English Language Learners). Moreover, the results about the comparisons among students from different school settings are based on small sample sizes (43 urban school students and 80 suburban school students), and therefore those results cannot be generalized. Second, quantitative results suggest that two items are sensitive to gender and SES. However, examination of the content of the items did not provide enough information for us to identify possible causes. Further research such as interviews with students about their understanding and perspectives about those two items are needed. Third, this study only provided preliminary evidence for the LP levels. Further validation of the LP will be conducted in a future study. More specifically, we will revise the assessments and use the revised assessments to collect data to further validate the LP.

Acknowledgements This work was supported by Education Testing Service CBAL (Cognitively Based Assessment of, for, and as Learning) Initiative. The authors wish to thank Randy Bennett for his valuable support throughout the research.

References

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*, 389–421.

Ben-Zvi Assaraf, O., & Orion, N. (2005). Development of system thinking skills in the context of earth systems education. *Journal of Research in Science Teaching, 42*, 400–419.

Ben-Zvi Assaraf, O., & Orion, N. (2010). Systems thinking skills at the elementary school level. *Journal of Research in Science Teaching, 5*, 540–563.

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education, 94*(5), 765–793.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: SAGE Publications.

Chi, M. T. H., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science, 36*(1), 1–61.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Dauer, J. T., Momsen, J. L., Speth, E. B., Makohon-Moore, S. C., & Long, T. M. (2013). Analyzing change in students' gene-to-evolution models in college-level introductory biology. *Journal of Research in Science Teaching, 50*, 639–659.

Duncan, R. G., Rogat, A. D., & Yarden, A. (2009). A learning progression for deepening students' understanding of modern genetics across the 5th–10th grades. *Journal of Research in Science Teaching, 46*, 655–674.

Forbes, C. T., Zangori, L., & Schwartz, C. V. (2015). Empirical validation of integrated learning performances for hydrologic phenomena: 3rd-grade students' model-driven explanation-construction. *Journal of Research in Science Teaching, 52*, 895–921.

Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching, 49*, 1181–1210.

Furtak, E. M., & Heredia, S. C. (2014). Exploring the influence of learning progressions in two teacher communities. *Journal of Research in Science Teaching, 51*, 982–1020.

Gotwals, A. W., & Songer, N. B. (2013). Validity evidence for learning progression-based assessment items that fuse core disciplinary ideas and science practices. *Journal of Research in Science Teaching, 50*(5), 597–626.

Gunckel, K., Covitt, B., Salinas, I., & Anderson, C. W. (2012). A learning progression for water in socio-ecological systems. *Journal of Research in Science Teaching, 49*(7), 843–868.

Hadenfeldt, J. C., Neumann, K., Bernholt, S., Liu, X., & Parchmann, I. (2016). Students' progression in understanding the matter concept. *Journal of Research in Science Teaching, 53*, 683–708.

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert-novice understanding of complex systems. *Journal of the Learning Sciences, 16*, 307–331.

Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science, 28*, 127–138.

Hogan, K. (2000). Assessing students' systems reasoning in ecology. *Journal of Biological Education, 35*, 22–28.

Hokayem, H., & Gotwals, A. W. (2016). Early elementary students' understanding of complex ecosystems: A learning progression approach. *Journal of Research in Science Teaching, 53*(10), 1524–1545.

Hokayem, H., Ma, J., & Jin, H. (2015). A learning progression for feedback loop reasoning at lower elementary level. *Journal of Biological Education. 49*(3), 246–260.

Holland, W., & Wainer, H. (1993). *Differential item functioning*. New York, NY: Routledge.

Jin, H., & Anderson, C. W. (2012a). Development of assessments for a learning progression on carbon cycling in socio-ecological systems. In A. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 151–182). Rotterdam, The Netherlands: Sense Publishers.

Jin, H., & Anderson, C. W. (2012b) A learning progression for energy in socio-ecological systems. *Journal of Research in Science Teaching, 49*(9), 1149–1180.

Jin, H., Johnson, M. E., Shin, H.-J., & Anderson, C. W. (2017). Promoting student progressions in science classrooms: A video study. *Journal of Research in Science Teaching*, *54*(7), 852–883.

Jin, H., Mikeska, J., Hokayem, H., & Mavronikolas, E. (2017, April). *Learning progression research: Toward the coherence in teaching and learning of science*. Paper presented at the annual conference of the National Association for Research in Science Teaching (NARST). San Antonio, TX: NARST.

Jin, H., Shin, H.-J., Johnson, E. M., Kim, J., & Anderson, C. W. (2015). Developing learning progression-based teacher knowledge measures. *Journal of Research in Science Teaching, 52*(9), 1269–1295.

Jin, H., & Wei, X. (2014). Using ideas from the history of science and linguistics to develop a learning progression for energy in socio-ecological systems. In R. F. Chen, A. Eisenkraft, F. Fortus, J. Krajcik, K. Neumann, J. C. Nordine, & A. Scheff (Eds.), *Teaching and learning of energy in K-12 Education* (pp. 157–174). New York, NY: Springer.

Jin, H., Zhan, L., & Anderson, C. W. (2013). Developing a fine-grained learning progression framework for carbon-transforming processes. *International Journal of Science Education*, *35*(10), 1663–1697.

Johnson, P., & Tymms, P. (2011). The emergence of a learning progression in middle school chemistry. *Journal of Research in Science Teaching, 48*, 849–877.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). New York, NY: American Council on Education, Macmillan.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lee, O., & Luykx, A. (2007). Science education and student diversity: Race/ethnicity, language, culture, and socioeconomic status. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 171–198). New York, NY: Routledge.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*, 675–698.

Neumann, K., Viering, T., Boone, W., & Fischer, H. E. (2013). Towards a learning progression of energy. *Journal of Research in Science Teaching, 50*(2), 162–188.

NGSS Lead States (2013). *Next generation science standards: For states, by states*. Washington, D.C.: Achieve, Inc..

National Research Council (2007). *Taking science to school: Learning and teaching science in grade K-8*. Washington, D.C.: The National Academies Press.

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

Plummer, J. D., & Krajcik, J. (2010). Building a learning progression for celestial motion: Elementary levels from an earth-based perspective. *Journal of Research in Science Teaching, 47*, 76–787.

Plummer, J. D., & Maynard, L. (2014). Building a learning progression for celestial motion: An exploration of students' reasoning about the seasons. *Journal of Research in Science Teaching, 51*, 902–929.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Resnick, M. (1996). Beyond the centralized mindset. *The Journal of the Learning Sciences, 5*(1), 1–22.

Resnick, M., & Wilensky, U. (1998). Diving into complexity: Developing probabilistic decentralized thinking through role-playing activities. *The Journal of the Learning Sciences, 7*(2), 153–172.

Ripple, W. J., Larsen, E. J., Renkin, R. A., & Smith, D. W. (2001). Trophic cascades among wolves, elk, and aspen on Yellowstone National Park's northern range. *Biological Conservation, 102*, 227–334.

Robitzsch, A., Kiefer, T., & Wu, M. (2017). *TAM: Test analysis modules*. R package version 1.99992–0. Retrieved from https://CRAN.R-project.org/package=TAM

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Boston, MA: Kluwer.

Schwarz, C. V., Reiser, B., Davis, E. A., Kenyon, L., Acher, A., Fortus, D., Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632–654.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 14*(1–2), 1–98.

Sudweeks, R. R., & Tolman, R. R. (1993). Empirical versus subjective procedures for identifying gender differences in science test items. *Journal of Research in Science Teaching, 30*(1), 3–19.

Sweeney, L. B., & Sterman, J. D. (2000). Bathtub dynamics: Initial results of a systems thinking inventory. *System Dynamics Review, 16*, 249–286.

Sweeney, L. B., & Sterman, J. D. (2007). Thinking about systems: Student and teacher conceptions of natural and social systems. *System Dynamics Review, 23*, 285–312.

Wilensky, U., & Resnick, M. (1999). Thinking in levels: A dynamic systems approach to making sense of the world. *Journal of Science Education and Technology, 8*, 3–19.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*, 716–730.