A Validation Framework for Science Learning Progression Research

Hui Jin

Student and Teacher Research Centre, Educational Testing Service, Princeton, United States

Peter van Rijn

ETS Global, Amsterdam, the Netherlands

John C. Moore

Department of Ecosystem Science and Sustainability, and Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, United States

Malcolm I. Bauer

Cognitive Sciences, Educational Testing Service, Princeton, United States

Yamina Pressler

Graduate Degree Program in Ecology, and Natural Resource Ecology Laboratory. Colorado State University, Fort Collins, United States

Nissa Yestness

Department of Ecosystem Science and Sustainability, and Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, United States

Correspondence: Hui Jin; (609) 734-5742; hjin@ets.org

Jin, H., van Rijn, P., Moore, J. C., Bauer, M. I., Pressler, Y., & Yestness, N. (2019). A validation framework for science learning progression research. *International Journal of Science Education*, 1-23. doi:10.1080/09500693.2019.1606471

Abstract

This article provides a validation framework for research on the development and use of science Learning Progressions (LPs). The framework describes how evidence from various sources can be used to establish an interpretive argument and a validity argument at five stages of LP research—development, scoring, generalization, extrapolation, and use. The interpretation argument contains the interpretation (i.e., the LP and conclusions about students' proficiency generated based on the LP) and the use of the LP. The validity argument specifies how the evidence from various sources supports the interpretation and the use of the LP. Examples from our prior and current research are used to illustrate the validation activities and analyses that can be conducted at each of the five stages. When conducting an LP study, researchers may use one or more validation activities or analyses that are theoretically necessary and practically applicable in their specific research contexts.

Keywords: assessment, learning progression, validation

Educators, administrators, and policymakers interpret test scores to assess learning outcomes and performance, to formulate policy, and to take action (Messick, 1989).

Validation is a process of evaluating the extent to which the proposed interpretations and the uses of test scores are plausible and appropriate (Kane, 2006). Traditional assessment and validation standards have been tested with the development of learning progressions (LPs) – “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” (National Research Council [NRC], 2007, p. 219).” LPs are cognitive models developed based on the researchers’ interpretation of students’ responses in interviews and written assessments. They have been used to guide the development and revision of a coordinated set of components, including student assessment, curriculum, classroom teaching strategies, teacher professional development materials, and teacher knowledge measures, We argue that LPs as well as the interpretations generated based on the LPs (e.g., students’ proficiency) must be validated before they are used to inform curriculum, instruction, and professional development programs.

LP researchers have long recognized the critical role of validation in LP research (Anderson, 2008; Duncan & Hmelo-Silver, 2009). Several validation strategies are often employed in the LP research. The design-based research method is used to revise and refine LPs in iterative cycles (e.g., Anderson et al., 2018; Breslyn, McGinnis, McDonald, & Hestness, 2016). Item Response Theory (IRT) analysis and the associated Wright Maps are used to show whether the LP levels are differentiated from each other and in the right order (e.g., Herrmann-Abell & DeBoer, 2018; Neumann, Viering, Boone, & Fischer, 2013; Rivet & Kastens, 2012). Some studies articulate an interpretive argument and a validity argument about the LP (e.g., Gotwals & Songer, 2013).

The above studies provide examples for using individual validation strategies in LP research. However, a comprehensive overview of validation of LP is needed. Based on the existing research base and our prior research, we developed a framework that shows when and how different activities and analyses can be used to validate the LP, and how the interpretive argument and validity argument can be generated in this process. Examples from our prior and current research are used to illustrate those validation activities and analyses. When conducting an LP study, the researchers may use one or more validation activities or analyses that are theoretically necessary and practically applicable in the specific research contexts.

Background

Researchers have developed LPs for different scientific ideas and scientific practices in recent years. Many of these LPs described levels of understanding with an upper anchor, a lower anchor, and intermediate levels connecting these two anchors. To provide a fine-grained description of student learning, some researchers identify multiple progress variables of an LP, with each progress variable elaborating the development in one dimension of the LP (Wilson, 2009).

Duncan and Hmelo-Silver (2009) provide a normative description of the LP structure:

> "First, LPs are focused on a few foundational and generative disciplinary ideas and practices (akin to the progress variables of the Bear Assessment System [BAS]; Wilson, 2009). … Second, these progressions are bounded by an upper anchor describing what students are expected to know and be able to do by the end of the progression; this anchor is informed by analyses of the domain as well as societal expectations. They are also bounded by a lower anchor describing the developers' assumptions about the prior knowledge and skills of learners as they enter the progression. Third, LPs describe varying levels of achievement as the intermediate steps between the two anchors. (pp. 606-607)"

Given the controversy over the LP approaches, we first situate our research in the recent debate on LPs that focuses on two closely related but distinct views of students' intuitive ideas. The "knowledge-in-pieces" view treats students' ideas as weakly connected pieces of knowledge, i.e., p-prims and facets; and claims that different ideas may be activated in different contexts (diSessa, 1993; Minstrell, 2000). The "knowledge-as-theory" view sees students' ideas as coherent and theory-like; and claims that students use naïve theories to explain phenomena across contexts (e.g., Carey & Spelke, 1994; Vosniadou, Vamvakoussi, & Skopeliti, 2008). The initiators of the LP approaches mostly drew upon the ideas from "knowledge-as-theory" view in developing LPs (NRC, 2007; Smith, Wiser, Anderson, & Krajcik, 2006). A major critique of LP research, coming from the "knowledge-in-pieces" tenet, is that LPs tend to present development as linear and sequential, and such presentations do not capture the dynamics and contextual factors of learning (Hammer & Sikorski, 2015). Regarding this critique, we provide two points of view.

First, student learning can be analyzed from different perspectives. On the one hand, science learning is complex and dynamic, and therefore it is meaningful to investigate the fragmentation and contextuality of learning. On the other hand, there are salient trends and patterns in student learning and development. LPs that capture those trends and patterns have important implications for classroom teaching (Duncan & Rivet, 2013; Duschl, Maeng, & Sezen, 2011). Both perspectives are powerful in developing cognitive models applicable to science classrooms, but neither of them alone provides an encyclopedic description of science learning. LPs are just one of many different types of cognitive models. They by no means should be viewed as authoritative depictions of science learning.

Second, in a systematic review of LP research (Jin, Mikeska, Hokayem, & Mavronikolas, April 2017), we found that many researchers present LPs in terms of several achievement levels (Hokeyam & Gotwals, 2016; Lehrer & Schauble, 2012; Rivet & Kastens, 2012; Schwarz et al., 2009). In these LPs, each achievement level describes a reasoning pattern or a broad concept. A few researchers developed LPs to capture the dynamics and contextual factors of learning (e.g., Johnson

& Tymms, 2011; Stevens, Delgado, & Krajcik, 2010). For example, Johnson and Tymms (2011) developed an LP for the concept of substance. The LP contains 52 ideas organized on a map in terms of difficulty ranges and conceptual categories. These ideas are pieces of knowledge because many of them are context-specific, and the researchers do not use a broad reasoning pattern or concept to connect related ideas. As an example, if we were to assess student reasoning about a burning candle, three ideas emerge—the candle decreases in mass, a candle flame produces new water, a candle produces new carbon dioxide in the LP map. These context-specific ideas fall into different conceptual categories and have different difficulty ranges on the map. They are not connected by a broad reasoning pattern or concept. In brief, as the aforementioned example illustrates, although we found many LPs focus on the salient patterns of student thinking, there are LPs that capture the fragmentation and dynamics of learning.

In our research, we developed LPs that take the form of sequential achievement levels. Note that our research by no means covers all perspectives within the LP research or represents a generalized LP approach. In this article, we use the work in two strands as examples to explain how to conduct the activities suggested in the validation framework. In one strand, we use the LP approaches to study the ways of thinking that students use to understand the dynamic interactions within and across social-ecological systems (ecosystems, atmosphere, and human social economic systems). We have developed two LPs in this strand. One LP describes how students from upper elementary to high schools use matter and energy as a conceptual tool to analyze phenomena related to the carbon cycle (Jin & Anderson, 2012; Mohan, Chen, & Anderson, 2009). The second LP focuses on secondary students' understanding of the interdependent relationships among organisms in ecosystems and humans' impacts on those interactions (Jin, Shin, Hokayem, Qureshi, & Jenkins, 2017). Together, these two LPs describes student development in using systems thinking to understand the complex social-ecological systems (see Moore et al. 2015; Moore & de Ruiter, 2012 for the science about systems thinking). In the second strand, we are developing an LP for quantification in two NGSS disciplinary core ideas: energy in physical sciences and ecosystems in life sciences. More specifically, we define quantification as the ability to analyze phenomena through identifying and conceptualizing measurable variables and understanding the relationship among variables.

The Validation Framework

A systematic validation must generate two forms of argument – an interpretive argument and a validity argument. Kane (2006, p. 23) explains that the interpretive argument "…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" , while the validity argument "…provides an evaluation of the interpretive argument." Moreover, validation should be conducted throughout the research, from assessment development to the interpretation and the use of assessment scores (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Based on these ideas and our prior research, we developed a framework that guides validation throughout the LP research (Figure 1).

[Insert Figure 1 about Here]

Within our framework, an interpretive argument (shaded rectangle) and a validity argument (dotted-line rectangle) are established through validation at five stages—development, scoring, generalization, extrapolation, and use. At the first four stages, interpretations, including the LP and conclusions generated based on the LP (e.g., students' proficiency), are constructed through a chain of inference: inferring LP and LP-based scores from responses; inferring the proficiency of students from observed scores; inferring the proficiency of students in a broader domain from the proficiency of students on the assessed construct. These inferences are made based on six assumptions about students' understanding and learning (Assumptions 1-6; Figure 1 and discussed below). The last stage is the use of the LP in schools and professional development programs. The use of the LP is evaluated by two arguments (Arguments 7 and 8). At each stage of the LP research, evidence from different sources is collected to evaluate the relevant assumption(s). When the evidence refutes the assumptions, it provides useful information for revising the LP, the items, and the scoring rubrics. When the evidence supports the assumptions, a validity argument can be established.

The Development Stage

At the development stage, the researchers define the assessment construct—the concepts, principles, ideas, and practices that the LP is about; they also develop items and tasks to assess the construct. Two assumptions at the development stage are listed below. Assumption 1 is about defining and specifying the construct, while Assumption 2 is about designing assessment for the construct.

- Assumption 1: The assessment construct (i.e., the upper anchor of the LP) addresses important ideas, concepts, and principles in science curriculum.
- Assumption 2: The items are effective in diagnosing the reasoning of students, including students who have not received formal instruction on the relevant knowledge.

A common approach of validating these two assumptions is to consult experts. In this approach, it is important to include experts with different expertise (e.g., scientists, science education researchers, and science teachers) so that the evaluation will cover content accuracy, cognitive concerns, and pedagogical concerns. Documentation that keeps track of evaluations and revisions can be used to provide evidence regarding how validity is enhanced through constant revision of the construct and the assessment tasks.

The evaluation of Assumption 2 must consider a unique feature of LPs—LPs describe student development over “a broad span of time”. This feature requires that the LP-associated assessment must diagnose the thinking of students across a wide age range and varying achievement levels (Jin & Anderson, 2012b). More importantly, the items and tasks must effectively elicit the intuitive ideas from younger students and students who have not received formal instruction of relevant instruction. At the minimum, such effective items and tasks use language and scenarios that make sense to those students.

Assumption 2 can be evaluated using the triangulation of multiple data sources. Although written assessments are useful in collecting data from a large population and form the basis of the quantitative analysis, the written responses provided by students may not provide enough details for in-depth interpretation. Therefore, think-aloud interviews and clinical interviews are often used to validate Assumption 2. Think-aloud interview data provide information about the *response process*—the cognitive process that a student uses to generate responses for a written item or task. In think-aloud interviews, students are instructed to talk out loud about what they think as they work through a written item or task, often with prompts such as, "talk more" and "say aloud whatever you are thinking" (Ericsson & Simon, 1980). In clinical interviews (*sensu* Piaget, 1929, Russ, Lee, & Sherin, 2012), *probing questions* are used to capture the knowledge and ways of thinking and reasoning that the students use to interpret a phenomenon or solve a problem (Chi, 1997). Unlike think-aloud interviews, clinical interviews target the product rather than the process of performing a task.

In our ongoing research on LP for quantification, we used both think-aloud interviews and clinical interviews with high school students to evaluate how well the assessment items assess the construct—identifying variables and understanding the relationships among variables. Below we present a think-aloud excerpt and a clinical-interview excerpt about a scenario: comparing the cold water in a bathtub with the hot water in a teacup (Figure 2). The purpose of the interviews is to assess how students identify and distinguish between extensive and intensive variables. Temperature is an intensive variable that does not depend on the amount of water. Thermal energy is an extensive variable that depends on the amount of water. Therefore, while the hot water in the teacup has a higher temperature, the cold water in the bathtub has more thermal energy.

[Insert Figure 2 about Here]

Think-aloud Excerpt:

Student A: Think about these two containers of water. On the left, we have a bathtub of cold water, and on the right, we have a cup of hot water. Please compare the cold water in the bathtub with the hot water in the cup. Do you think the water in the bathtub and the water in the cup have energy? Please explain your answer. I don’t think either has energy because if they are resting in the bathtub and the cup, they're not going to have any potential because they're not raised at all. And since the water isn’t moving, there is no kinetic energy.

Clinical-interview Excerpt:

Interviewer: So have you learned about thermal energy?

Student B: Yes.

Interviewer: What does it mean?

Student B: Thermal energy is energy due to heat. It’s energy represented by heat. So you have like mechanical energy which we totally talked about with physics and then this energy doesn't take place with, like, the motion I guess.

Interviewer: Okay. So which one of those vessels do you think has more thermal energy?

Student B: Um, I would say the water or that’s in the cup, not the bathtub. I kind of forget, umm…

Interviewer: Well, what are you thinking? What are you going through?

Student B: Um, like thermal energy is like related to the temperature. So it makes sense that if you have like two things of the same size, the hotter one would have a greater amount of thermal energy. But depending on like the size of the containers, if you have like something really large size, only like one degree less, that will obviously have more energy in it related to heat.

The think-aloud excerpt suggests that Student A is using kinetic energy and gravitational potential energy to analyze the scenario. The student provided a correct answer about the energy of the two objects (a teacup of hot water and a bathtub of cold water). He explained that, since the two objects are not at a position above the ground and have no movement, they do not have energy. Although the item was designed to assess student ability to identify and distinguish between temperature and thermal energy, the data of Student A's response process suggests that the student drew upon other knowledge and he analyzed the scenario correctly. In the clinical interview excerpt, the interviewer used probing questions to elicit Student B's thinking. Student B's responses to those questions suggest that thermal energy is an unfamiliar term to Student B. At first, Student B provided an incorrect answer that the water in the cup had more thermal energy. However, Student B's responses toward the end of this episode suggested that he clearly identified two different variables in this situation—an intensive variable not depending on the size and an extensive variable depending on the size. This evidence indicates that, although Student B were able to differentiate intensive and extensive variables in the specific context, his unfamiliarity with the term "thermal energy" affects how well he understood the question and explained his thinking to the reviewer.

The data from both interviews uncovered that the use of the term thermal energy is not an indicator of the quantification ability. Instead, students' unfamiliarity with the term affected how well the student respond to the questions. Therefore, we revised the item to assess quantification rather than assessing whether the students are able to name the scientific terms. Instead of asking students to compare temperature and thermal energy, the revised item (Figure 3) asks students why the hot water in a teacup and the cold water in a bathtub cause different results. In order to explain this, students must distinguish between intensive and extensive variables. While the intensive quantity (temperature) explains why hot water burns people, the extensive quantity (thermal energy) explains why the cold water makes an ice cube melt. More specifically, the hot water in the teacup has a very high temperature, meaning that the water particles move very fast. The fast moving water particles interact with the particles of the skin cells and accelerate their motion, causing the cells to break up. While the cold water in the bathtub has a lower temperature, there are a lot of cold water. Therefore, the cold water in the bathtub provides enough energy to melt a big ice cube.

[Insert Figure 3 about Here]

Four responses are provided below. Both Response A and Response B provide a clear distinction between temperature and energy/heat and use these two differentiated variables to explain why the cold water in the bathtub is more likely to make a large ice cube to melt completely. None of these responses uses the term ‘thermal energy’. Unlike Responses A and B., Responses C and D use an undifferentiated variable to explain why the cold water in the bathtub is more likely to melt the ice cube. Response C uses the variable of ‘coldness’/‘hotness’ (i.e., being cold or hot). Response D uses the variable of temperature. Although these two responses use different terms, they suggest the same understanding—a single variable that conflates the meaning of temperature and the meaning of energy is used to explain the phenomenon.

Response A: In the tea cup, although much hotter, has much less water in the cup. Once an ice cube is placed, it will rip through much of it while rapidly bringing the temperature of the teacup water down to a point where the water has much less energy to completely melt it, therefore the icecube [ice cube] won't completely melt the large ice cube. In the bathtub, although much cooler, has much more water at a constant temperature that can melt ice. The ice can float around to different parts of the 25 C [°C] water and take energy from different points to completely melt it. One ice cube wont dramatically affect the bathtubs overall temperature, while one ice cube will dramatically affect the teacups temperature.

Response B: The hot water in the tea cup [teacup] is not enough to melt completely the large cube of ice. Eventually the heat exchanged will make all the water too cold to keep melting the ice cube. Instead, despite the fact that it has a lower temperature, the water in the tub is in a much larger amount, and the heat exchanged with the cube is not enough to significantly affect its temperature.

Response C: the [The] cold water in the bathtub is not cold enough to keep the ice freeze but the water in the cup is hot enough to burn someone.

Response D: The hot water is more likely to burn someone because of its high temperature [temperature], well the cold water makes an ice cube melt because water can break things down/dissolve stuff.

The above example shows how clinical interviews and think-aloud interviews are used to evaluate Assumption 2. The interview data provide rich information about students' response process and their understanding of the questions. Such information allows researchers to evaluate whether the assessment items and tasks are effective in eliciting students' thinking (Assumption 2).

The Scoring Stage

At the scoring stage, researchers collect assessment data from students, interpret the data to identify patterns of student thinking and reasoning. They use these patterns to develop and revise the achievement levels of the LP. The LP-based rubrics are developed and used to score students' responses for the LP levels. In this process, researchers infer scores from student responses. The validity of *the inference from responses to scores* is based on Assumption 3.

- Assumption 3: The LP and the LP-based scoring rubrics capture salient patterns of students' reasoning and present the development in a meaningful way.

In developing the LP, a common approach is using qualitative methods (e.g., thematic analysis and constant comparative method) to identify students' reasoning patterns inductively and then use these patterns to build the LP levels. Assumption 3 is about the meaning of the LP levels. The LP levels must tell a coherent story of student development that is compatible with the major findings and theories in cognition and learning sciences (Anderson, 2008). As such, theories and findings from literature can be used as validity evidence for the existence of the levels and the developmental trend described in the levels (Assumption 3).

Regarding Assumption 3, a critical issue that LP researchers often encounter is to find out whether two or more ideas from students should be categorized into the same level. In our research, we frequently faced this issue when using the general LP levels to develop specific scoring rubrics for individual items. Apparently, it would be difficult to find evidence from literature to validate this particular aspect of Assumption 3. Here, we use an example from our research on LP for systems thinking to show how clinical interview data can be used to assist decision-making on this issue. The example uses the Yellowstone National Park item (See Figure 4),

[Insert Figure 4 about Here]

The item assesses middle and high school students' understanding of one aspect of the interdependent relationships in ecosystems—the trophic cascade. It seeks an explicit mechanism to explain how plants changed when humans removed the top predators from the Yellowstone ecosystem. It is used to differentiate two LP levels. At Level 1, students only recognize direct and immediate relations among organisms. Some examples of Level 1 responses are: “The feces and dead bodies of the wolves were fertilizers for the plants to live and grow.” “The wolves once provided carbon dioxide for the plants to live and grow.” “The wolves once spread seeds that helped the plants reproduce.” At Level 2, students recognize that all populations in an ecosystem are connected in food chains/webs; they are able to identify distant relations among populations. Two examples of Level 2 responses are: “They might have decreased because if the wolves die whatever the wolves were eating increased so the animal the wolves were eating probaly [probably] ate plants so if those animals were eating those plants then the plants would all die because that animal ate them all.” “The disapperance [disappearance] of wolves follows the increase of mice, rabbits, ect., which are herbivores. This stands to 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.”

While analyzing the written data, we found two student ideas that may or may not belong to the same level. One idea is that killing wolves and the disappearance of vegetation are connected because wolves’ dead bodies or feces provide nutrients and fertilizer for plants to grow. Written Response 1 is an example of this idea. The other idea is that killing wolves and the disappearance of vegetation are connected because both wolves and plants are in a system and all species in a system depend on each for the system to persist. Written Response 2 is an example of this idea.

> Written Response 1: The wolf might have been fertilaizing [fertilizing] the area around the trees and now the trees have nothing making the land fertal [fertile].
>
> Written Response 2: The disappearance of the wolf population may have caused the decrease of plant populations because the wolves kept the ecosystem in balance.

Clearly, the first idea should be categorized as Level 1 because the student did not use the idea of a food chain to connect the top predators (wolves) and plants. We were struggling with whether the second idea belonged to Level 1 or Level 2 of the scoring rubrics. On the one hand, the student recognized that all components in a system were connected and affected each other. On the other hand, the student did not apply that general understanding to the specific context—how the wolves and plants in the Yellowstone National Park were connected. It seems that this idea can be categorized to either Level 1 or Level 2. It also seems reasonable to create a new level between existing Level 1 and Level 2 to capture this idea. Our clinical interview data provided a solution to this dilemma. Below is an excerpt from an interview about the event happened in the Yellowstone National Park:

[The student read a card that has pictures and text illustrating the context.]

Interviewer: So do you think there could be some connection between killing the wolves and the disappearance of the trees?

Student: Yes.

Interviewer: What is the connection?

Student: I think the connection is that once a wolf dies, it starts to decompose and once it decomposes it provides the nutrients for the soil to have for there to be growth in the trees; and it supplies the proper nutrients and everything the tree needs to grow. So I think that is why once the wolves died out there weren't any natural nutrients in the soil and that is why the soil was desolate: now barren. So that is why the plants started dying out.

Interviewer: So scientists actually found there were other animals that provided nutrients. If there were other animals, do you think it still is the same reason?

Student: So scientists also found that there are still other animals and the plants still died out without the wolves?

Interviewer: Yeah.

Student: I think they might have had a system there where everything depended on each other, the same as the lynx and the hare. So without the other, they needed when … they both need each other in order to thrive. So I think without that one component, just threw everything into chaos and that is why everything started disrupting that environment: the Yellowstone National Park.

In this interview excerpt, the student first explained that killing wolves and the disappearance of vegetation were connected because wolves provided nutrients for trees to grow. This explanation suggested the first idea discussed above. To further probe the student's thinking, the interviewer told the student that other animals could provide nutrients for the trees. The student then responded that wolves and plants must be connected because they were in the same system, but the student could not explain that connection in terms of the food chain (wolves-herbivores-plants). This response indicates the second idea discussed above. In brief, the student used two ideas simultaneously to explain the Yellowstone task. Although the student understood that all components in a system must be connected, the student did not demonstrate the ability to apply that understanding to the specific context—wolves and plants in the Yellowstone ecosystem. Therefore, we categorized both ideas as Level 1 indicators. In this case, the rich information collected via clinical interview was used to validate the scoring rubrics for the written responses. When scoring the responses for an item, the data may indicate that a new level need be created, or two levels need to be collapsed into one, or one level need to be split into two. In the same indication appears across many items, the revision of the scoring rubric may lead to the revision of the LP. Shea and Duncan (2013) provide detailed descriptions of the revisions of the LP levels.

The Generalization Stage

At the generalization stage, researchers perform quantitative analyses to the score to infer the proficiency of students from observed scores. LPs are models that describe students' development. Therefore, the LP levels should be differentiated from each other, showing that the levels actually exist. It is also important that the order of the LP levels should not only make conceptual sense but also have empirical evidence. As such, Assumption 4, as elaborated below, is about the differentiation among and the order of the LP levels.

- Assumption 4: The administered assessment items yield sufficient evidence for the LP levels, including the evidence that the levels are differentiated from each other and the evidence of the order of the levels.

Various quantitative techniques can be used to evaluate this assumption. For example, a widely used technique is to apply IRT analysis to the test scores and use the Wright Map to present the analysis results. The Wright Map shows whether the achievement levels are differentiated from each other and whether the order of the achievement levels makes sense (e.g., Steedle & Shavelson, 2009; Wilson, 2012). Here, we propose a validation approach that can be used with items that differentiate responses among some but not all levels of an LP. In this approach, an LP is conceptualized as an underlying continuous variable (for more details, see Graf & van Rijn, 2016; van Rijn, Graf, & Deane, 2014). As an illustration, we analyzed data (*n* = 596 students) from an assessment on the LP for interdependent relationships in ecosystems. The assessment consisted of two forms, one for middle-school students (*n* = 298) and one for high-school students (*n* = 298) with a total of 13 unique items. The middle-school form had 10 items, the high-school form had 11 items, and both forms had 8 common items. The items targeted either levels 1 and 2 or levels 1, 2, and 3 of the LP and were directly scored with respect to these levels. Table 1 presents the levels of the LP.

Unintelligible responses, irrelevant responses, and ‘I don’t know’ type responses were scored 0. In such responses, students do not describe any relationships among organisms in an ecosystem.

[Insert Table 1 about Here]

We performed the quantitative analyses for all students as one group. We fit a partial credit model (PCM; Masters, 1982), which is an extension of the Rasch model to polytomous items, and a constrained partial credit model (CPCM). We made use of the item response theory software developed by Haberman (2013), which employs marginal maximum likelihood estimation of item parameters. Both models assume a latent variable θ and a mathematical function that links the probabilities of item scores to θ and a set of item parameters. The item parameters of the PCM locate the transitions between score categories (which are linked to LP levels) on the scale of the latent variable θ. This link to LP levels provides a rationale for the CPCM in which the item parameters that are related to the same LP level are constrained to be equal across items. The main parameters of this model can then be interpreted as the locations of the transitions between LP levels on the latent scale. Furthermore, an estimate of these parameters and θ can be used directly to locate a student’s standing on the LP. For more details and an illustration of this modeling approach, see Appendix A.

The results of fitting the models are shown in Table 2. The Bayesian Information Criterion (BIC) of the PCM is lower than that of the CPCM, indicating a better relative fit for the PCM. However, the parameters of the CPCM are ordered in the expected direction (i.e., higher LP levels are more difficult than lower LP levels) and can therefore be interpreted as average LP level transitions. We will discuss some further results using the CPCM for illustrative purposes.

[Insert Table 2 about Here]

The left panel of Figure 5 shows the estimated functions for item category probabilities based on the CPCM. These functions apply to all items, although they have to be renormalized for items that target only Level 1 and Level 2 so that the probabilities sum up to one. In the figure, the curves for different levels are differentiated from each other, indicating the existence of the levels. The vertical dashed lines indicate the transitions between two adjacent LP levels. Such a transition is the point where the probabilities of the adjacent LP levels are equal. For example, the green line indicates the transition between Level 1 and Level 2, and is the point where the red dashed curve (Level 1) and the green dashed curve (Level 2) intersect. The model does not specify that these transitions are ordered, so the ordering that is found here is an empirical finding that supports the theoretical ordering of the LP levels. In other words, the empirical data indicate that the transition between Level 2 and Level 3 (the blue vertical line) appears after the transition between Level 1 and Level 2 (the green vertical line), which appears after the transition between reference (score 0) and Level 1 (the red vertical line). This order is the same as the order of the LP levels that were determined conceptually. Furthermore, these transitions define latent intervals that are linked to the LP levels because the items were directly scored with respect to LP levels. This is seen in the right panel of Figure 5, which depicts the frequency distribution of estimated θ and the LP level transitions. As such, Figure 5 provides a piece of validity evidence that supports Assumption 4—the administered assessment items yield evidence for the LP levels.

[Insert Figure 5 about Here]

In the paragraphs above, we have discussed how to validate the assumption about the LP levels. A question that is frequently raised at conferences and workshops deals with whether we make further inferences about the level of each student. Although inferences about an individual student’s level could provide useful information for the design of curriculum and instruction, controversy exists regarding whether such inferences can be made. If only quantitative methodology is considered, two methods can be used to classify students onto LP levels, with each method based on an assumption about the coherence of the students' ideas.

First, each student can be mapped onto a single LP level, assuming that students tend to use reasoning at one level to construct explanations across items. Research in the learning sciences provides a way for us to decide whether this method is appropriate. Two major theories about the coherence of student thinking are the knowledge-as-theory (Gopnik & Wellman, 1994) and the knowledge-in-pieces theory (diSessa, 1993). Both theories are supported by evidence from empirical research. Some researchers found that students' ideas, although less coherent than scientists' ideas, are theory-like (Reiner, Slotta, Chi, & Resnick, 2000; Vosniadou, et al., 2008). Other researchers found that students tend to hold many idiosyncratic and fragmented ideas, and those ideas can be at different levels of proficiency (diSessa, 1993; Minstrell & Stimpson, 1996). Apparently, neither of these theories supports the assumption that student ideas are extremely coherent—using reasoning at one level to consistently reason about a variety of phenomena. Most students in our study provided responses at different levels for different assessment items, providing additional evidence to refute the assumption that students tend to use reasoning at a single level to construct explanations.

Second, instead of assuming that students apply one reasoning pattern across contexts, we can use quantitative analysis to examine how coherent students' ideas are. For each student, we can compute the possibilities of the student's responses at different levels of an LP. If the result shows evidence of 'knowledge-in-pieces' (e.g., probabilities of responses at different levels are about the same), the statement that students' ideas are coherent to a certain degree does not hold and we cannot make any inferences about the students' levels. In such a situation, Assumption 5, as listed below, is not meaningful. Otherwise, Assumption 5 needs to be evaluated.

- Assumption 5: Classifying students onto levels is appropriately determined.

As an example, we analyzed the above data on students' understanding of interdependent relationships in ecosystems (n = 596 students). The analysis produced evidence of the coherence of students' ideas and therefore supported Assumption 5. Under Assumption 5, we then infer the levels of individual students. In the paragraphs that follow, we elaborate on this process. A benefit of the above CPCM (Table 2) is that if the model holds, then there is a direct link between the LP level, the ability estimate, and the total score. Since two forms were used, a basic equating table is easily produced by linking total score ranges for each form to LP levels, as shown in Table 3.

[Insert Table 3 about Here]

With Table 3, we can compute level classification probabilities for each student by using multiple imputations (i.e., repeated sampling from the posterior of the ability and using the transitions as cut-offs). These classification probabilities give an indication of the consistency of the student responses and can be useful for reporting. Some examples are shown in Table 4. The level classification probabilities are based on 10,000 samples from the individual posterior distribution. It can be seen that each student produced responses at adjacent levels, suggesting that those students' ideas are coherent to a certain degree. Therefore, Table 4 provides the evidence for the assumption that students' ideas are to a certain degree coherent (Assumption 5). Based on this assumption, we can make inferences about the levels of individual students. For example, we can make the following inferences about the levels of the students listed in Table 4:

- Student 200 sometimes identifies direct relationships among organisms (e.g., rabbits eat grass).
- Student 201 and Student 220 both rely on Level 1 and Level 2 reasoning. Student 201 is able to identify direct relationships among organisms and begin to pay attention to distant relationships (e.g., top predators are connected to plants) and patterns (e.g., predator-prey

cycle) in ecosystems. Student 220 is able to identify direct relationships in ecosystems, and sometimes identify distant relationships and patterns in ecosystems.

- Student 210 is able to identify direct relationships, indirect relationships, and patterns in ecosystems. Student 210 also begins to reason about the mechanism (e.g., energy pyramid and feedback loops) behind those relationships and patterns.

[Insert Table 4 about Here]

The Extrapolation Stage

At the extrapolation stage, the interpretation of student proficiency extends into a much larger domain. In the case of LP research, extrapolation is about extending student proficiency measured by the LP into the domain of a science discipline or all science disciplines. This inference is based on Assumption 6.

- Assumption 6: The proficiency demonstrated in the assessments in related to the students' proficiency in other relevant courses.

Validity evidence that is based on relations to other variables allows researchers to evaluate whether the intended construct is related to other variables measuring similar constructs or related criteria (AERA, et al., 2014) and therefore can be used to validate Assumption 6. For example, correlation measures can be computed between the LP scores and external indicators (e.g., students' scores in other science courses, teacher evaluations).

The Use Stage

Kane (2013) urges researchers to take into account three score-based decisions. *Intended effect* refers to the extent to which the intended outcomes are achieved. *Adverse impact* refers to the differential impact on groups, particularly adverse impact on legally protected groups. *Unintended* *systematic effect* refers to positive and negative systematic effects, particularly in education. Here, we discuss how to validate two assumptions related to the intended effect (Assumption 7) and adverse effect (Assumption 8).

- Assumption 7: The LP is useful for teachers to help students move towards the upper anchor of the LP.
- Assumption 8: The LP and associated materials, when used appropriately by teachers, do not having an adverse impact on identifiable subgroups of students.

LPs are cognitive models developed based on a broad body of literature, including science education, the history and philosophy of science, and the learning sciences. They are also validated based on empirical data from real students. As such, LPs have the potential to promote teaching and learning of science in schools. However, empirical evidence suggests that many teachers have difficulty in developing the upper anchor understanding, in using assessment to elicit students' ideas at different LP levels, and in designing learning activities to target those intuitive ideas (Authors, 2015, 2017; Furtak, 2012; Furtak & Heredia, 2014; Furtak, Morrison, & Kroog, 2014). Therefore, practicality is an important quality of LP and associated components (e.g., curriculum and professional development materials). Assumption 7 is about the practicality of LP. In the paragraphs that follow, we first discuss strategies to enhance the practicality of LPs. Then, we suggest approaches to evaluate the practicality of the LP.

Three strategies can be used to enhance the practicality of LPs. First, researchers may develop LPs that present the developmental trend of students concisely; they may also clarify the distinction between the developmental trend and lesson sequences to teachers. Researchers found that LPs with too many details overwhelm teachers, especially beginning teachers who have little classroom teaching experience (Furtak, Thompson, Braaten, & Windschitl, 2012). Therefore, it is very important that an LP focus on the main developmental trends rather than including every detail of students' understanding and development. In our ongoing research, we also found that teachers tend to conflate an LP with a lesson sequence. A common misconception of LPs is that the LP levels describe learning goals for a sequence of lessons or learning activities. Therefore, it is important to explain the distinction between LPs and lesson sequences to teachers. LPs focus on the 'ways of knowing', while traditional lesson sequences often describe the 'content' that accompanies the knowing—a sequence from less sophisticated concepts and principles to more sophisticated ones. Many LPs describe students' naïve ideas and alternative conceptions at the lower anchor and the intermediate levels. These ideas cannot be used as learning goals, but they provide useful information for developing learning activities that target those naïve ideas and use those naïve ideas as founds of knowledge. Recent studies investigate how teachers understand and use LP as a model of student development (Gunckel, Covitt, & Salinas, 2018; von Aufschnaiter & Alonzo, 2018).

Second, LPs should have accompanying instructional components that provide educative supports for teachers (Furtak et al., 2012). Empirical LP studies indicate that teachers in particular need support with two components of Pedagogical Content Knowledge (PCK), understanding student thinking and deciding on next instructional moves (Furtak, 2012; Jin et al., 2015a; Jin, Johnson, Shin, & Anderson, 2017). Therefore, educative supports on these two PCK components should be provided to supplement the LPs. Here, we use the Plant Unit developed by the Environmental Literacy project as an example (See Schramm et al., 2012; Project websites: http://envlit.educ.msu.edu/; http://www.pathwaysproject.kbs.msu.edu/; http://carbontime.bscs.org/home). The unit uses stories of typical students to help teachers understand student thinking and make decisions on instruction. At the beginning of the Plant Unit, a driving question about plant growth is provided with three stories. The driving question is: *Little acorns can grow into big, heavy oak trees. Where does all the mass of an oak tree come from?* Next, stories of three students, Adrienne (a typical Level 2 student), Beatrice (a typical Level 3 student), and Carla (a typical Level 4 student) are provided. Each story describes a student's explanation of the driving question. Throughout the Plant Unit, "the story of Adrienne checkpoint" is provided to explain the alternative conceptions and learning difficulties of a typical level 2 student and how the learning activities were designed in ways to help Adrienne be aware of her own thinking and master scientific mechanistic reasoning gradually.

Third, professional development programs may involve teachers in the development of LPs. Various professional development models such as Professional Learning Community (Furtak & Heredia, 2014; Richmond & Manokore, 2011) can be used to design research activities that involve teachers in the development of LPs and associated materials. For example, Furtak and Heredia (2014) compared two teacher communities, including one where teachers co-developed an LP for natural selection with the researchers and one where teachers did not. Teachers in both communities used the LP to design and teach lessons. The researchers found that teachers involved in LP development used the LP in meaningful ways, while teachers in the other community struggled to make sense of the LP within the accountability context of their schools. Teacher involvement may take different forms and be at various levels. Teachers may provide the researchers with resources such as assessment tasks and ideas about student intuition and learning difficulties. They may work with the researchers on co-developing LPs. They may also test the LP in their classrooms and provide feedback to the researchers. Researchers can determine the appropriate level of teacher involvement based on project goals and the participating teachers' knowledge level.

To evaluate the effectiveness of the above strategies as well as other strategies that are aimed at enhancing the practicality of the LP (Assumption 7), evidence of teaching practice need to be obtained. Lesson videos, classroom observations, lesson plans, and teacher interviews provide rich information for researchers to investigate how and how much teachers understand and use the LP and associated materials in their day-to-day teaching. In each of these cases, the data are used to make inferences about the degree to which the LP and associated materials are useful and educative for teachers. Assumption 7 can also be evaluated by student learning outcomes. For example, the researcher can carry out an intervention study in which teachers use the LP and associated materials in their instruction. Learning gains of students taught by teachers using the LP can be compared to results from students who were not taught in this manner to evaluate to what extent the LP-based intervention produced improved learning outcomes.

Assumption 8 refers to the LP and associated materials, when used appropriately by teachers, not having an adverse impact on identifiable subgroups of students. Differential Item Functioning (DIF) analysis can be used to flag assessment items that might be biased against certain subgroups such as females, students with relatively low socioeconomic status, or English as a second language students. With a large sampling of schools, where students are representatively sampled, observational, interview, and outcome data can be used to make inferences about the degree students are engaged in productive learning strategies, and if the use of LPs differentially affected those groups in negative ways.

Conclusion

This article provides a framework that describes how a variety of activities and analyses can be used to validate the LP, the interpretations based on the LP, and the subsequent uses of the LP. In the framework, an interpretive argument and a validity argument are established through validation at five stages—development, scoring, generalization, extrapolation, and use. The interpretation argument contains the LP, the interpretation about student performance based on the LP, and the ideas about how the LP should be used in schools. The validity argument specifies how various pieces of evidence support these interpretations and use. When using this framework, researchers choose the validation activities and analyses that are theoretically necessary and practically applicable for their specific research contexts. We present empirical evidence to support the development through extrapolation stages and assumptions 1 through 6. A major limitation of this article is that, since our research is still ongoing, we have not obtained empirical data that evaluate the effectiveness and usefulness of the LPs for teachers (Assumption 7) and the impact on certain subgroups of students (Assumption 8).

Acknowledgement

This work was supported by Institute of Education Sciences under Grant R305A160219, National Science Foundation under Grant DUE 0832173, and National Aeronautics and Space Administration under Grant NNX10AT51A, and Education Testing Service CBAL (Cognitively Based Assessment of, for, and as Learning) Initiative. The authors wish to thank Michael Kane, Caroline Wylie, Jessica Andrews, and James Carlson at Educational Testing Service for their insightful comments and suggestions.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, C. W. (2008). *Conceptual and empirical validation of LPs*. East Lansing, MI: Michigan State University.

Anderson, C. W., de los Santos, E. X., Bodbyl, S., Covitt, B. A., Edwards, K. D., Hancock, I., James Brian, . . . Welch, M. M. (2018). Designing educational systems to support enactment of the Next Generation Science Standards. *Journal of Research in Science Teaching*, *55*, 1026-1052.

Breslyn, W., McGinnis, J. R., McDonald, R. C., & Hestness, E. (2016). Developing a learning progression for sea level rise, a major impact of climate change. *Journal of Research in Science Teaching, 53*(10), 1471-1499. doi:10.1002/tea.21333

Carey, S., & Spelke, E. (1994). Domain-specific knowledge and conceptual change. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture.* Cambridge; New York: Cambridge University Press.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 6*(3), 271-315.

diSessa, A. (1993). Towards an epistemology of physics. *Cognition and Instruction, 10*(2-3), 105-225.

Duncan, R. G., & Hmelo-Silver, C. (2009). LPs: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching, 46*, 606-609.

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*(6), 655-674.

Duschl, R. A. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education, 32*, 268-291.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data.*Psychological Review, 87*(3), 215.

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 LPs in two teacher communities. *Journal of Research in Science Teaching, 51*, 982-1020.

Furtak, E. M., Thompson, J., Braaten, M., & Windschitl, M. (2012). LPs to support ambitious teaching practices. In A. Alonzo & A. W. Gotwals (Eds.), *LPs in science: Current challenges and future directions* (pp. 461-472). Rotterdam, The Netherlands: Sense Publishers.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*. New York: Cambridge University Press.

Gotwals, A. W., & Alonzo, A. C. (2012). Introduction: Leaping into LPs in science. In A. Alonzo & A. W. Gotwals (Eds.), *LPs in science: Current challenges and future directions* (pp. 3-12). The Netherlands: Sense Publishers.

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

Graf, E. A., & van Rijn, P.W. (2016). LPs as a guide for design: Recommendations based on observations from a mathematics assessment. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 165-189). New York, NY: Routledge.

Gunckel, K. L., Covitt, B. A., & Salinas, I. (2018). Learning progressions as tools for supporting teacher content knowledge and pedagogical content knowledge about water in environmental systems. *Journal of Research in Science Teaching, 55*, 1339-1362.

Hammer, D., & Sikorski, T-R. (2015). Implications of complexity for research on learning progressions. *Science Education*, 99, 424–431.

Herrmann-Abell, C. F., & DeBoer, G. E. (2018). Investigating a learning progression for energy ideas from upper elementary through high school. *Journal of Research in Science Teaching, 55,* 68-93.

Hokayem, Hayat, & Gotwals, Amelia Wenk. (2016). Early elementary students’ understanding of.complex ecosystems: A learning progression approach. *Journal of Research in Science Teaching*, 53, 1524–1545.

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 progression in science classrooms: A video study. *Journal of Research in Science Teaching*. doi: 10.1002/tea.21388

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.

Jin, H., Shin, H.-J., Hokayem, H., Qureshi, F., & Jenkins, T. (2017). Secondary students' understanding of ecosystems: A learning progression approach. *International Journal of Science and Mathematics Education*. doi:10.1007/s10763-017-9864-9

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*, 1269-1295.

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 & Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1-73.

Lehrer, R., & Schauble, L. (2012). Seeding evolutionary thinking by engaging children in modeling its foundations. *Science Education*, 96(4), 701–724.

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

Messick, S. (2016). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.

Minstrell, J. (2000). Student thinking and related assessment: Creating a facet-based learning environment. In Committee on the Evaluation of National and State Assessments of Educational Progress, N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. MItchell, & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 44-73). Washington, DC: National Academy Press.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year LP for carbon cycling in socio-ecological systems. *Journal of Research in Science Teaching, 46*(6), 675-698.

Moore, J. C., Anderson, C. W., Berkowitz, A., Covitt, B. C., Gunckel, K., Hartley, L… Yestness, N. (2015, April). *Learning pathways to environmental science literacy*. Paper presented at the meeting of the National Association for Research in Science Teaching, Chicago, IL.

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

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

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.

The NGSS Lead States. (2013b). *Next generation science standards: For states, by states*. Washington DC: Achieve.

Piaget, J. (1929). *The child's conception of the world*. London, UK: Routledge & Kegan Paul.

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*(7), 768-787.

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

Richmond, G., & Manokore, V. (2011). Identifying elements critical for functional and sustainable professional learning communities. *Science Education*, *95*(3), 543-570.

Rivet, A. E., & Kastens, K. A. (2012). Developing a construct-based assessment to examine students' analogical reasoning around physical models in earth science. *Journal of Research in Science Teaching, 49*, 713-743.

Russ, R. S., Lee, V. R., & Sherin, B. L. (2012). Framing in cognitive clinical interviews about intuitive science knowledge: Dynamic student understandings of the discourse interaction. *Science Education*, *96*, 573-599.

Schramm, J., Keeling, E., Figueroa, D., Mohan, L., Johnson, E. M., & Anderson, C. W. (2012). *Plant growth and gas exchange*. *Culturally relevant ecology, LPs and environmental literacy*, Fort Collins, CO: Colorado State University.

Schwarz, C. V., Reiser, B., Davis, E. A., Kenyon, L., Achér, 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.

Shea, N. A., & Duncan, R. G. (2013). From theory to data: The process of refining LPs. *The Journal of the Learning Sciences, 22*(1), 7-32.

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, 14*(1&2).

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of LP level diagnoses. *Journal of Research in Science Teaching, 46,* 699-715.

Stevens, S. Y., Delgado, C., & Krajcik, J. (2010). Developing a hypothetical multi-dimensional LP for the nature of matter. *Journal of Environmental Education, 47*(6), 687-715.

Stevens, S. Y., Gotwals, A. W., Jin, H., & Barrett, J. (2015). Learning progressions research planning and design. In M. Solem, N. Huynh, & D. Boehm (Eds.), *Learning progressions for maps, geospatial technology, and spatial thinking: A research handbook* (pp. 22-43). Newcastle, UK: Cambridge Scholars Publishing.

von Aufschnaiter, C., & Alonzo, A. C. (2018). Foundations of formative assessment: Introducing a learning progression to guide preservice physics teachers' video-based interpretation of student thinking. *Applied Measurement in Education, 31*(2), 113-127.

Vosniadou, S., Vamvakoussi, X., & Skopeliti, I. (2008). The framework theory approach to the problem of conceptual change. In S. Vosniadou (Ed.), *International Handbook of Research on Conceptual Change* (pp. 3-34). New York: Routledge.

Table 1.

The Learning Progression for Interdependent Relationships in Ecosystems

Learning Progression Level	Level Description
0. Reference	The student does not describe any relationships among organisms (e.g., I don’t know).
1. Individual Organisms	The student describes relationships in terms of needs of individual organisms or random causes
2. Relationships and Patterns	The student identifies distant relations and patterns of interactions in ecosystems, but cannot use systems thinking concepts to successfully explain those patterns.
3: Mechanisms	The student uses systems thinking concepts (exponential growth and/or carrying capacity; energy pyramid; feedback loop) to construct a causal mechanism that explains phenomena about interactions in ecosystems.

Table 2.

Results of Fitting Two Models

Model	Parameters	Log-Likelihood	BIC	Reliability
PCM	33	-4676.5	9564	.881
CPCM	4	-5040.7	10107	.875

Table 3

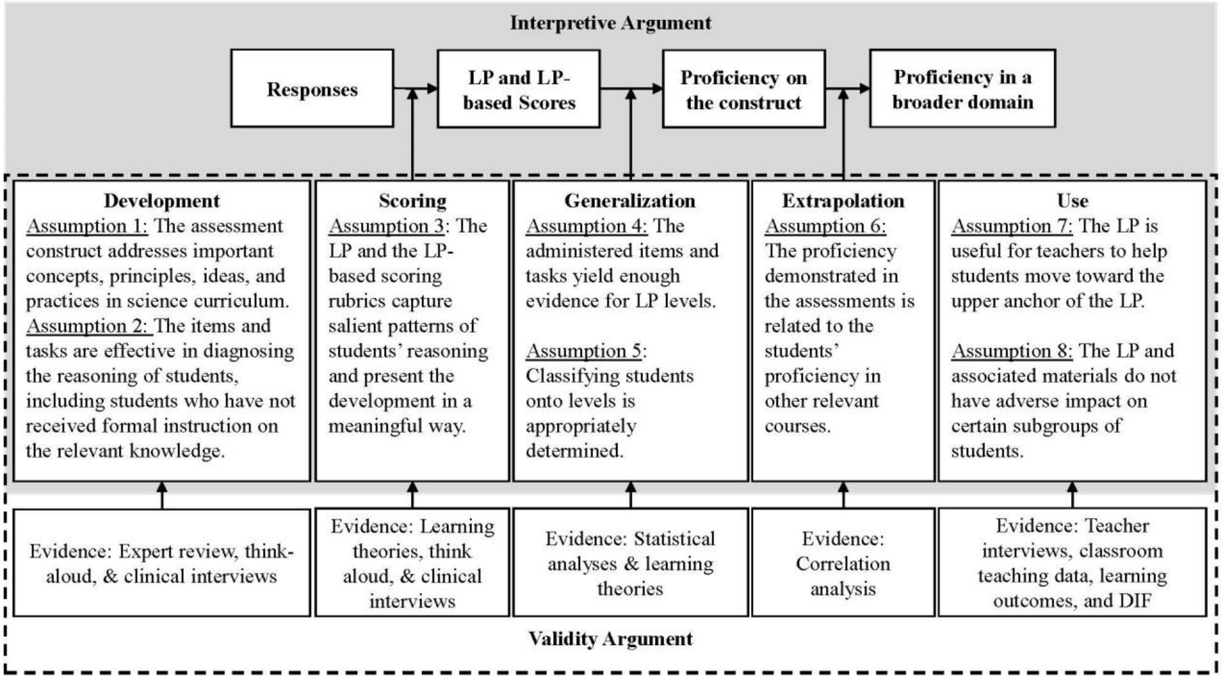
Equating Table for Relating Total Scores on Two Forms to Learning Progression Levels

Learning Progression Levels	Total Score Range (Middle-School Form)	Total Score Range (High-School Form)
0 (Reference)	0 – 6	0 – 5
Level 1	7 – 15	6 – 16
Level 2	16 – 20	17 – 23
Level 3	21	24 – 25

Table 4

Examples of Level Classification Probabilities for Four Students

Student	LP Level Classification Probability (based on posterior)			
	0 (Reference)	Level 1	Level 2	Level 3
200	.27	.73	.00	.00
201	.00	.89	.11	00
210	.00	.00	.68	.32
220	.00	.50	.50	.00



Assumption 1: The assessment construct addresses important concepts, principles, ideas, and practices in science curriculum.

Assumption 2: The items and tasks are effective in diagnosing the reasoning of students, including students who have not received formal instruction on the relevant knowledge.

Assumption 3: The LP and the LP-based scoring rubrics capture salient patterns of students' reasoning and present the development in a meaningful way.

Assumption 4: The administered items and tasks yield enough evidence for LP levels.

Assumption 5: Classifying students onto levels is appropriately determined.

Assumption 6: The proficiency demonstrated in the assessments is related to the students' proficiency in other relevant courses.

Assumption 7: The LP is useful for teachers to help students move toward the upper anchor of the LP.

Assumption 8: The LP and associated materials do not have adverse impact on certain subgroups of students.

Evidence: Expert review, think-aloud, & clinical interviews

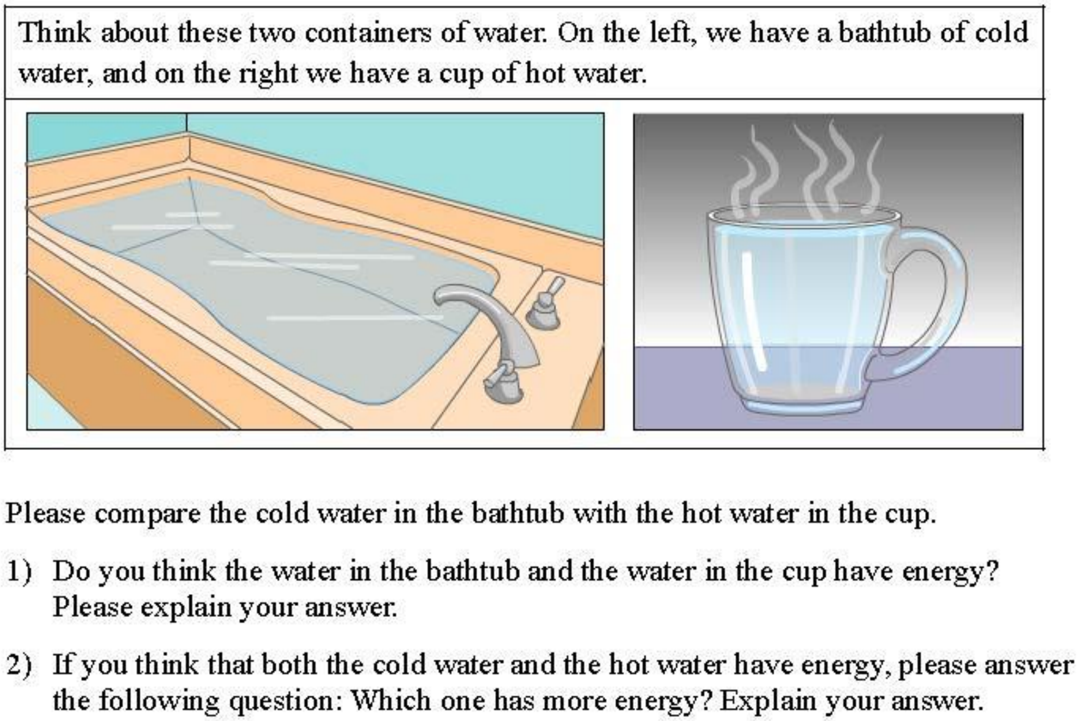
Evidence: Learning theories, think aloud, & clinical interviews

Evidence: Statistical analyses & learning theories

Evidence: Correlation analysis

Evidence: Teacher interviews, classroom teaching data, learning outcomes, and DIF

Figure 1. A Validation Framework for Science LPs.



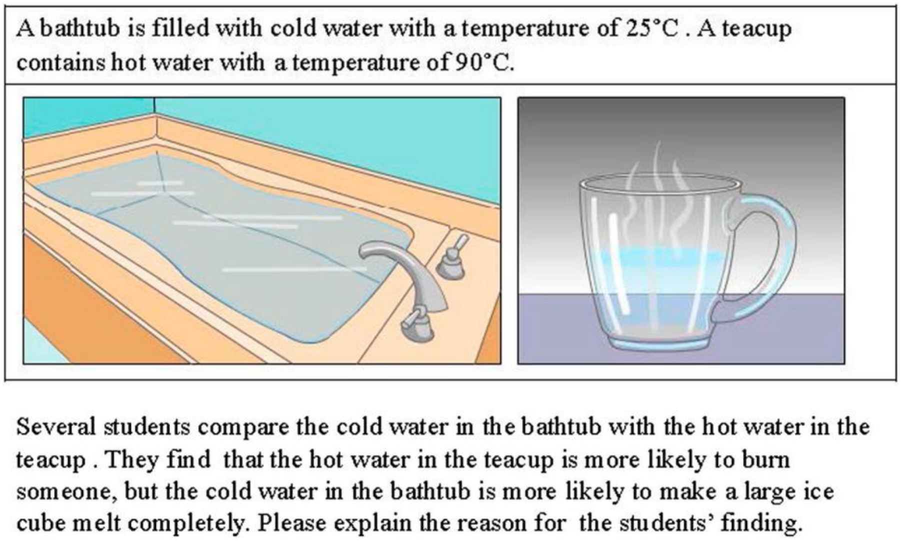
Think about these two containers of water. On the left, we have a bathtub of cold water, and on the right we have a cup of hot water.

Please compare the cold water in the bathtub with the hot water in the cup.

1) Do you think the water in the bathtub and the water in the cup have energy? Please explain your answer.

2) If you think that both the cold water and the hot water have energy, please answer the following question: Which one has more energy? Explain your answer.

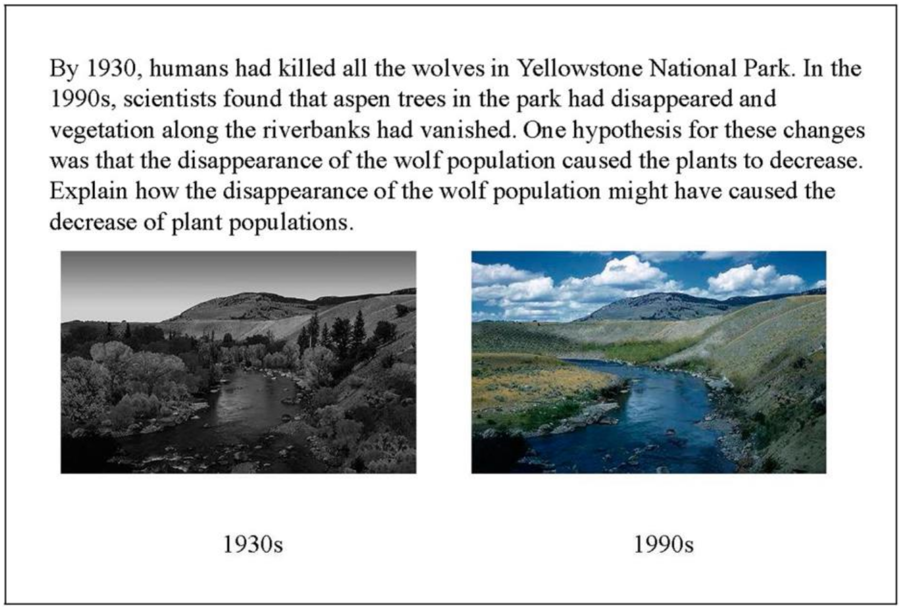
Figure 2. The initial version of the Bathtub and Teacup Item.



A bathtub is filled with cold water with a temperature of 25°C . A teacup contains hot water with a temperature of 90°C.

Several students compare the cold water in the bathtub with the hot water in the teacup . They find that the hot water in the teacup is more likely to burn someone, but the cold water in the bathtub is more likely to make a large ice cube melt completely. Please explain the reason for the students' finding.

Figure 3. The revised version of the Bathtub and Teacup Item.



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

Figure 4. The Yellowstone National Park Item.

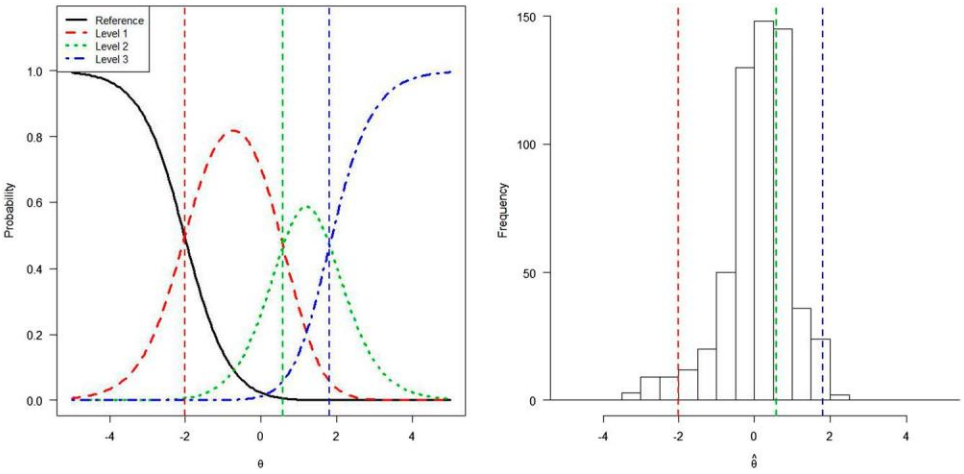


Figure 5. Estimated item category response functions of CPCM (left) and frequency distribution of estimated (right). The horizontal dashed lines indicate the transitions between adjacent learning progression levels.

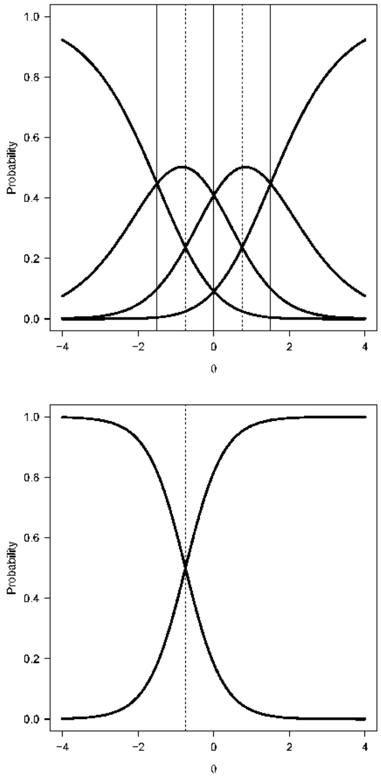
Appendix A. Description of PCM and CPCM

In the partial credit model (PCM; Masters, 1982), the probability of a score k on item j is modeled as follows:

$$P\left(Y_{ij} = k\right) = \frac{\exp\left(\sum_{h=0}^{k}\left[\theta_i - \eta_{ih}\right]\right)}{\sum_{v=0}^{m_j}\exp\left(\sum_{h=0}^{v}\left[\theta_i - \eta_{ih}\right]\right)}, \quad k = 0,1,\mathrm{K}\ ,m_j.$$

In this equation, the parameters η_{ih} are threshold parameters, indicating the point at which the item probabilities of category k and $k-1$ are equal, and θ_i is the latent variable of person i. As noted, if item responses are scored directly with respect to LP levels, the PCM can be constrained such that the threshold parameters that relate to the same LP level are equal across items. This holds even if not all items target the same LP levels, but different subsets. This constrained PCM model is referred to as the CPCM.

Figure A1 shows the item response functions for three CPCM items for an LP with three levels and parameters $\boldsymbol{\eta} = (-1.5, 0, 1.5)$. The item for which the IRF is displayed in the top of Figure A1 addresses LP levels 1, 2, and 3; the item in the middle addresses level 2 only; the item in the bottom addresses levels 1 and 3. The vertical solid lines indicate transitions between adjacent LP levels and the vertical dashed lines indicate transitions between non-adjacent LP levels. The thresholds parameters create intervals of θ that are associated with LP levels. It can be seen that these intervals are the same across items. Finally, note that the PCM would have six thresholds parameters, compared to only three for the CPCM.



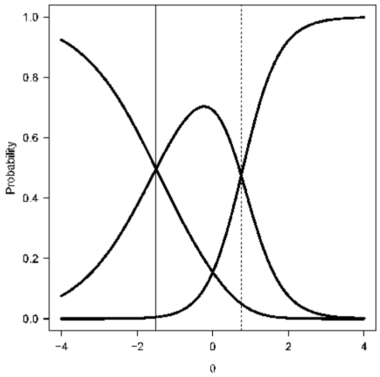


Figure A1. Examples of item response functions under CPCM.