

*Research Article***Validity Evidence for Learning Progression-Based Assessment Items That Fuse Core Disciplinary Ideas and Science Practices**Amelia Wenk Gotwals¹ and Nancy Butler Songer²¹*Teacher Education, Michigan State University, Erickson Hall, East Lansing, Michigan*²*School of Education, University of Michigan, Ann Arbor, Michigan**Received 11 January 2012; Accepted 25 January 2013*

Abstract: This article evaluates a validity argument for the degree to which assessment tasks are able to provide evidence about knowledge that fuses information from a progression of core disciplinary ideas in ecology and a progression for the scientific practice of developing evidence-based explanations. The article describes the interpretive framework for the argument, including evidence for how well the assessment tasks are matched to the learning progressions and the methods for interpreting students' responses to the tasks. Findings from a dual-pronged validity study that includes a think-aloud analysis and an item difficulty analysis are presented as evidence. The findings suggest that the tasks provide opportunities for students at multiple ability levels to show evidence of both successes and struggles with the development of knowledge that fuses core disciplinary ideas with the scientific practice of developing evidence-based explanations. In addition, these tasks are generally able to distinguish between different ability-level students. However, some of the assumptions in the interpretive argument are not supported, such as the inability of the data to provide evidence that might neatly place students at a given level on our progressions. Implications for the assessment system, specifically, how responses are elicited from students, are discussed. In addition, we discuss the implications of our findings for defining and redesigning learning progressions. © 2013 Wiley Periodicals, Inc. *J Res Sci Teach* 50: 597–626, 2013

Keywords: learning progressions; assessment; validity

Teaching students to become scientifically literate citizens, who are able to make informed decisions about pressing scientific issues, entails more than asking students to memorize facts. Rather, students must be engaged in key scientific practices around core disciplinary ideas. In order to move toward this goal, Achieve will deliver the Next Generation Science Standards (NGSS) in 2013 that were created in partnership with the National Academies of Science's National Research Council (NRC), National Science Teachers Association (NSTA), and the American Association for the Advancement of Science (AAAS). The precursor document titled, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts and Core Ideas* (NRC, 2011), outlined a vision for science education that was informed by research in science education and the learning sciences, including an emphasis on learning progressions, a smaller number of core disciplinary ideas, and the integration of science practices with core disciplinary

Contract grant sponsor: National Science Foundation; Contract grant numbers: REC-0089283; REC-0129331;

Contract grant sponsor: Spencer Dissertation Fellowship.

Correspondence to: A. W. Gotwals; E-mail: gotwals@msu.edu

DOI 10.1002/tea.21083

Published online 28 February 2013 in Wiley Online Library (wileyonlinelibrary.com).

ideas as a means to deepen understandings of core ideas (NRC, 2007). Drawing from foundational theories of learning, the Framework (NRC, 2011) articulated three dimensions of science knowledge that should be emphasized within the NGSS and subsequently within standards, curricula, instruction, and assessment. The committee recommends that science education in grades K-12 be built around three major dimensions. These dimensions are:

- Scientific and engineering **practices**.
- **Crosscutting concepts** that unify the study of science and engineering through their common application across fields.
- **Core ideas in four disciplinary areas**: physical sciences; life sciences; earth and space sciences; and engineering, technology, and the applications of science (NRC, 2011; p. ES-1).

Recently, a handful of research groups have championed the view articulated in the Framework (NRC 2011) that there needs to be an explicit means of inextricably linking core disciplinary ideas with practices (e.g., Songer & Gotwals, 2012). This stance was adopted to bridge a gap introduced by prior work on inquiry science (e.g., American Association for the Advancement of Science [AAAS], 1993). While fostering knowledge that was a fusion of core disciplinary ideas with practices was promoted as desirable in these efforts, few standards, learning progressions, curricular units, or assessments of the 1990s and early 2000s provided resources that represented fused knowledge.

The Framework provides samples of performance expectations that are the “assessable” (NRC, 2011) version of the fused knowledge. Presumably the next step is to create and evaluate assessment tasks that are matched to these performance expectations. The performance expectations in the Framework (NRC, 2011) are modeled after those used by College Board (e.g., College Board, 2009), an organization that is also developing assessments that are matched to performance expectations. This article provides research illustrating the development and evaluation of resources to assess fused knowledge. We present and evaluate a validity argument (Kane, 2001) for assessment tasks that fuse information from a core disciplinary idea progression and scientific practice (evidence-based explanations) progression. The evaluation includes two types of validity evidence to illustrate the extent to which our assessment tasks were matched to our learning progressions focused on upper elementary students’ fused knowledge about ecology and explanations.

Why Learning Progressions?

Learning progressions are valuable as they provide “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. 214). Learning progressions also offer a promising framework for fusing core disciplinary ideas and practices into learning performances for students at multiple levels of sophistication (Corcoran, Mosher, & Rogat, 2009; Gotwals, Songer, & Bullard, 2012).

Designing assessments that provide evidence of student understandings at multiple points along a learning progression is challenging (Anderson, Alonzo, Smith, & Wilson, 2007). This challenge is even more pronounced when attempting to design and evaluate assessment tasks that focus on science knowledge that is a fusion of core disciplinary ideas and science practices (Gotwals et al., 2012). However, having an assessment system that is aligned with an underlying learning progression framework, contains assessment items that allow students at multiple levels to demonstrate their fused knowledge, and provides a means for interpreting students’ responses with respect to the learning progressions is essential if we are to make progress in research on learning progressions (Corcoran et al., 2009).

Assessments for Learning Progressions

Assessment includes the processes of gathering evidence about students' knowledge and abilities as related to the tasks to which they respond as well as making inferences from that evidence about what students know or can do more generally (Mislevy, Steinberg, & Almond, 2003; NRC, 2001). The design of complex assessments (like those needed to gather evidence about learning progressions) must "start around the inferences one wants to make, the observations one needs to ground them, the situations that will evoke those observations, and the chain of reasoning that connects them" (Messick, 1994, p. 17).

The Consortium for Policy Research in Education (CPRE) reports that the strength of learning progressions work relies on "assessments that measure student understanding of the key concepts or practices and can track their developmental progress over time" (Corcoran et al., 2009, p. 15). Assessments are essential for learning progressions because in order to understand the pathways that students take as they develop more sophisticated abilities and to gather validity evidence about the learning progressions themselves, we must have some way to measure and assess what students know and can do at multiple levels and over time. Research suggests that assessment instruments that are developed in coordination with learning progressions can provide more information about a larger range of students than typical assessments (Songer & Gotwals, 2012; Songer, Kelcey, & Gotwals, 2009) and offer more discriminatory power than traditional items (Liu, Li, Hofstедder, & Linn, 2008). In order to make strong arguments about the importance of learning progressions in guiding the development of assessment instruments, however, we need validity evidence that supports arguments that the assessment tasks do, indeed, provide information about the knowledge represented in the learning progressions.

We define learning progressions as one of several possible idealized sequences (Songer et al., 2009) that are also "partly hypothetical or inferential, since long-term longitudinal accounts of learning by individual students do not exist" (National Assessment Governing Board [NAGB], 2008, p. 90). In gathering validity evidence about learning progression-based assessments, the objective or construct must first be defined. In the case of learning progressions, this requires identification of the core disciplinary ideas and scientific practices represented at the multiple points along the learning progression. Upper anchors are often defined both by standards and societal expectations (Corcoran et al., 2009; NRC, 2007). For lower and middle levels, learning progressions may rely upon cognitive science research. However, this research is not complete (NAGB, 2008; NRC, 2007). While we know quite a bit about what young children understand in some content areas and contexts, we also recognize that learning does not always happen in a linear, stepwise fashion, thus making the articulation and empirical backing of one idealized learning path difficult to generate. It is possible that even the same learner might proceed differently through a knowledge development path in different contexts, so we recognize that while it might be possible to identify target upper and lower anchors of a progression, articulation of the intermediate learning points might be better described as "the messy middle" (Gotwals & Songer, 2010, p. 277). Compounding the challenge, we recognize that students do not always demonstrate consistent levels of understanding across core disciplinary ideas and contexts (Alonzo & Steedle, 2009; Steedle & Shavelson, 2009) and students often do not respond consistently to sets of items designed to tap the same underlying principles (e.g., Chi, Feltovich, & Glaser, 1981; Gotwals & Songer, 2010). Recognizing the tension that arises between the messiness of student learning and the need for greater systematicity in guiding learners towards more complex learning outcomes, we propose that learning progressions are valuable as templates for the systematic design of coordinated instructional, professional development, and assessment products (Songer et al., 2009).

Validity

The 1999 *Standards for Educational and Psychological Testing* define validity as “the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests” (AERA, APA, & NCME, 1999, p. 9). Using this definition, validity can be thought of as evidentiary reasoning (Mislevy, 2012) or as an argument structure, in which an interpretive argument is laid out, evidence is gathered, and the strength of the argument is evaluated (Kane, 1992). Developing an argument about an assessment must involve both a clear articulation of the intended knowledge and skills to be measured as well as matching this with empirical evidence of students interacting with the items or tasks. Kane (1992, 2001) differentiates between two connected arguments: (1) an interpretive argument, which lays out a framework for linking the inferences and assumptions about students’ responses on assessment tasks (scores) to the proposed interpretation and use of the scores, and (2) a validity argument, which uses empirical evidence to evaluate the “plausibility of the proposed interpretation by critically examining the inferences and assumptions in the interpretive argument” (Kane, 2001, p. 339).

This study focuses on evaluating an interpretive argument for the degree to which our assessment tasks are matched to knowledge that fuses information from our core disciplinary ideas and evidence-based explanation progressions. Evaluating the validity of an assessment for a given purpose is an iterative process. First, one must develop an interpretive argument based on the claims that one wants to make. Second, one must gather evidence relevant to the inferences and assumptions in the interpretive argument. Then this evidence must be evaluated, focusing especially on the most problematic parts of the interpretive argument (and adjusting the argument or inference or gathering more data, if necessary). Finally, this process must be repeated until the inferences in the interpretive argument are plausible or the interpretive argument is rejected (Kane, 2001, p. 330). However, validity is not an “all or nothing” concept.¹ Evidence is used to create an argument for the strength of making the proposed interpretations from students’ responses to assessment tasks.

In the following sections, we present our interpretive argument, in which we seek to examine how well our assessment tasks are able to generate evidence about the fused knowledge from our core disciplinary ideas and explanation progressions. To do this, we present our core disciplinary ideas and explanation progressions and assessment tasks and make inferences about how student responses to these tasks can be mapped onto the fusion of our core disciplinary ideas and explanation progressions. Then we present the evidence that we gathered and how we used this evidence to make a claim about how well our tasks provided evidence of students’ understandings based on our learning progression.

Mapping Student Responses to Learning Progressions: The Interpretive Argument

Kane (1992), states that, “inferences from test scores to theoretical constructs depend on assumptions included in the theory defining the construct” (p. 527). Thus, our first step is to define our learning progression and present a rationale for the importance of fusing core disciplinary ideas and practices. Below we present a description of how we use the products based on our learning progressions to represent the fusion of core ecological ideas and the practice of developing evidence-based explanations. The argument here is that these assessment tasks do, indeed, allow students at multiple levels the ability to illustrate their how they fuse core ecological ideas with the practice of developing explanations.

Fusing Knowledge About Ecology and Explanations

In order to have a detailed picture of students’ understandings in science, we must consider not only their understanding core disciplinary ideas, but also the ways in which students use these

ideas in order to interpret and explain scientific situations and phenomena. The Framework for Science Education Standards prioritizes learning goals that are a fusion of core disciplinary ideas with scientific practices (such as the construction of evidence-based explanations), called *performance expectations* (NRC, 2011). The Framework also states that learning science should feature "... a commitment to data and evidence as the foundation for developing claims. The argumentation and analysis that relate evidence and theory are also essential features of science; scientists need to be able to examine, review, and evaluate their own knowledge and ideas and critique those of others" (NRC, 2011, pp. 2–3). Developing a learning progression that articulates the ways in which students move from their everyday ways of arguing or explaining (e.g., Bricker & Bell, 2007) to our goals for developing coherent evidence-based explanations that incorporate claims, evidence, and reasoning, is an important step in understanding ways to better support students in developing this important practice.


In our work in fusing core disciplinary ideas with the practice of explanations, we began by developing two progressions, one for core disciplinary ideas (see Figure 1 for a simplified version) and one for explanations (Table 1). As described in Songer et al. (2009), our core disciplinary ideas progression was developed in conjunction with scientists, teachers, and educational researchers to represent a sequence of essential core disciplinary ideas for students to develop a more sophisticated ability to explain ecological phenomena. There are three strands of core ideas: classification, ecology, and biodiversity. However, in order to have students work meaningfully

	Classification Strand	Ecology Strand	Biodiversity Strand
6th Grade		<p>Complex Ecological Idea: A change in one species can affect different members of the food web...</p> <p>...</p> <p>Middle Ecological Idea: Plants and animals of a habitat can be connected in a food chain</p>	<p>Complex Biodiversity Idea: Humans and other factors affect biodiversity...</p> <p>...</p> <p>Middle Biodiversity Idea: Biodiversity differs in different areas...</p>
5th Grade	<p>Complex Classification Idea: Patterns of shared characteristics reveal the evolutionary history...</p> <p>...</p> <p>Middle Classification Idea: Organisms are grouped based on their structures...</p>		<p>Middle Biodiversity Idea: An area has a high biodiversity if it has both high richness and abundance</p> <p>...</p> <p>Basic Biodiversity Idea: A habitat is a place that provides food, water, shelter...</p>
4th Grade	<p>Middle Classification Idea: Organisms have different features that allow them to survive</p> <p>...</p> <p>Basic Classification Idea: There are observable features of living things</p>	<p>Middle Ecological Idea: Only a small fraction of energy at one level ... moves to the next level</p> <p>...</p> <p>Basic Ecological Idea: Every organism needs energy to live...</p>	

Figure 1. Modified core ideas progression.

Table 1

Practice progression for evidence-based explanations



7. Student is provided with a scientific question, and asked to construct a scientific explanation (including a claim, evidence and reasoning) (the process is not scaffolded)
6. Student is provided with a scientific question, and asked to construct a scientific explanation (including a claim, evidence and reasoning) (the process is scaffolded with hints about the core ideas)
5. Student is provided with a scientific question, and asked to construct a scientific explanation (including a claim, evidence and reasoning) (the process is scaffolded with hints about the core ideas and prompts for including the claim, evidence, and reasoning)
4. Student is provided with a scientific question, and asked to make a claim and back it up with evidence (the process is not scaffolded)
3. Student is provided with a scientific question, and asked to make a claim and back it up with evidence (the process is scaffolded with hints about the core ideas)
2. Student is provided with a scientific question, and asked to make a claim and back it up with evidence (the process is scaffolded with hints about the core ideas and prompts for including the claim and evidence)
1. Student is either provided with evidence and asked to choose the appropriate claim OR student is provided with a claim and asked to choose the appropriate evidence

with these core ideas, they must be joined with a practice from our evidence-based explanation progression (Table 1) into a performance expectation.

Similar to other researchers (e.g., McNeill, Lizotte, Krajcik, & Marx, 2006; Ruiz-Primo, Li, Tsai, & Schneider, 2010), our explanation progression focuses on three essential aspects of explanations around focal core disciplinary ideas: (1) articulation of *claims*; (2) use of appropriate and sufficient *evidence* to support these claims; and (3) use of *reasoning* that draws on scientific principles to explicitly link the evidence to the claim. Claims are assertions or conclusions in response to a scientific question (in our project, claims are either given to students by the teacher or curriculum or students create their own claims). Evidence consists of scientific data (either collected by students or given to students by the teacher or curriculum) used to support students' claims. Data used as evidence must be appropriate and sufficient. Appropriate data are relevant to the question or problem and support the claim. There is sufficient evidence when enough relevant data are used to convince someone of the accuracy of the claim (McNeill & Krajcik, 2007). Finally, reasoning is a justification that utilizes salient scientific principles to provide a bridge that associates the evidence in support of the claim. Our practice progression (Table 1) lays out the ways in which we scaffold students in developing more sophisticated abilities to engage with the practice of developing evidence-based explanations and the means for assessing this developing practice.

Pellegrino (2012), states that, “[o]ne benefit of carefully described learning progressions is that they can be used to guide the specification of learning performances. . . . The learning performance can in turn guide the development of tasks that allow one to observe and infer students’ levels of competence for major constructs. . . .” (p. 835). In our work, the fusion of core disciplinary ideas and practice occurs through performance expectations that are subsequently used to design each curricular activity and each assessment task. Based on earlier work (e.g., Songer et al., 2009) and in consultation with teachers and scientists, the project selected a specific core disciplinary idea and joined each one with a given level of the practice to create a fused product called the performance expectation.² Figure 2 presents an example of one idea in ecology taken from the core disciplinary ideas progression plus one level from our practice progression, and the performance expectation that is the fused product of their joining. Figure 2 also provides an example of an embedded assessment item matched to these same core disciplinary idea and

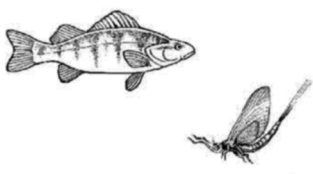
<p>Core Disciplinary Idea: Ecology 13: Because many animals rely on each other a change in the number of one species can affect different members of the web.</p>	<div style="text-align: right;"> Data Sheet 23 MS-ETS1 </div> <p>Scientific Question: How have recent changes in the Detroit River affected yellow perch populations?</p> <p>Write your scientific EXPLANATION:</p> <p><i>There have been a lot of yellow perch in the Detroit River. The yellow perch population because of the pollution in the water. The water is dirty. They also have dumped a lot of trash that has killed the fish. Over time there are many pollutants and too much fertilizer. There is a lot of algae in the water. The yellow perch have died. There are a lot of dead yellow perch in the water. The yellow perch population has decreased to 50% of what it was. After the pollution stopped in the Detroit River the yellow perch population increased to 75% of what it was.</i></p> <div style="border: 1px solid black; padding: 5px; width: fit-content;"> <p>Hint: Think about information from the story and the graphs to help you write your explanation.</p> <p>Hint: Think about what yellow perch eat.</p> <p>Hint: Think about what factors might affect the yellow perch's food.</p> </div> 
<p>Practice: Explanation 6: Student is provided with a scientific question and construct a scientific explanation (including a claim, evidence and reasoning) (the process is scaffolded with hints about the core ideas)</p>	
<p>Fused Core Idea-Practice Performance Expectation: Students construct a scientific explanation to address the question, how have recent changes in the Detroit River affected yellow perch populations?</p>	

Figure 2. Sample core ideas and science practice sections and their fusion into learning goals.

practice and a sample student response. Overall, the project team developed three curricular units of 8 weeks each and three summative assessments based on fused core disciplinary ideas and practices performance expectations.

Developing a Suite of Assessments Matched to the Fused Performance Expectations

While students may come into a science classroom able to construct some aspects of explanations or arguments (Berland & Reiser, 2010), writing coherent evidence-based explanations that include all of the essential aspects listed above (i.e., claims, evidence, and reasoning) is difficult for students (Gotwals & Songer, 2010; Songer & Gotwals, 2012; Songer et al., 2009; Ruiz-Primo et al., 2010; White & Frederiksen, 1998). Despite these struggles, with time, repeated exposures, and support students have been shown to make gains in using core disciplinary ideas to explain phenomena; students move from creating unsubstantiated or insufficiently substantiated claims to making claims that are backed by evidence and reasoning (Berland & McNeill, 2010; McNeill et al., 2006; Songer et al., 2009). Thus, our project has used educational scaffolds, structures that are placed strategically to help students better understand confusing or unfamiliar topics or to prompt them to utilize certain knowledge, as a way to support students in developing explanations around focal core disciplinary ideas and as a way to structure our assessment of their progressing abilities. Our work has shown that scaffolds are able to both improve the quality of evidence-based explanations as well as students’ abilities to integrate core ecological ideas and reasoning skills (see Table 1 and Songer et al., 2009; Songer & Gotwals, 2012).

Our work on assessment design has been documented in other places (e.g., Gotwals & Songer, 2010; Gotwals et al., 2012; Songer & Gotwals, 2012), so we will present a short description of our learning progression-based assessment design. In each assessment task, we fuse a core ecological idea from our core disciplinary idea progression (Figure 1) with a level of our practice progression (Table 1). To create the assessment task, we used written scaffolds, similar in structure to those in our curriculum that provide students with different levels of explanation tasks (Gotwals & Songer, 2010), making them close or proximal to our curriculum (Ruiz-Primo,

Shavelson, Hamilton, & Klein, 2002). For this analysis, our assessments included three task levels that varied as to the amount and type of scaffolding. In items at the minimal level, students are given evidence and are asked to choose a claim that matches the evidence (the first level in Table 1), in intermediate items, students construct an explanation with structural practice scaffolds (the fifth level in Table 1; that provide prompts and hints about the three components of evidence-based explanations), and in complex items, students construct an explanation with no scaffolds (the seventh level in Table 1). Figure 3 provides an example of an intermediate assessment item that includes explanation scaffolds (that prompt students to include a claim, evidence, and reasoning). The item in Figure 3 is designed to gather evidence about how students construct an explanation with core disciplinary ideas B7 in the core disciplinary ideas progression (i.e., “Biodiversity differs in different areas. It is a useful way of characterizing habitats, it tells you something about the quality of the habitat as a whole for a number of different organisms”). Figure 4 provides an example of a complex item that does not include any scaffolding. The item in Figure 4 is designed to gather evidence about how students construct an explanation with core disciplinary ideas C4 in the core disciplinary ideas progression (“Organisms (animals) have different features that they use to survive in different habitats. There are observable internal and external differences (some fly, some have scales, fur, wings, live in the water, etc.). Some of these differences are used to distinguish major groups”). For more details of the assessment system see Gotwals et al. (2012) or Songer and Gotwals (2012).

Empirical Evidence to Support Interpretations of Students' Learning Progression Levels

There should be multiple sources of evidence used in order to ensure that assessments are measuring their intended knowledge and skills (AERA, APA, & NCME, 1999). “The trustworthiness of the interpretation of the test should rest on empirical evidence that the assessments tasks tap the intended cognition” (NRC, 2001, p. 147). We conducted a series of empirical studies to obtain information about how well our assessment tasks were able to elicit student responses mapped to the fusion of knowledge from our core disciplinary ideas and explanation progressions. Specifically, we utilized a dual-pronged approach to gather empirical validity evidence: a think aloud and cognitive interview analysis and a difficulty analysis using Rasch modeling.

Data Collection

This study was conducted in the Detroit Public Schools (DPS), an urban district that has a total district enrollment of approximately 183,000 students in 263 schools. The project has a long history of working with students and teachers in DPS. DPS is characteristic of many urban school districts in the United States in that it contains a concentration of students of color, students from low-income families, and students learning English as a second language (e.g., 94% of DPS students characterize themselves as ethnic minorities and over 70% of students are eligible for free or reduced lunch).

We worked within the implementation of the curricular program in three schools' sixth grade classes. There was one teacher in each school, each of who had multiple sections of sixth grade science. The data for this study come from over 300 students' responses to a written assessment as well as responses from 20 low-, medium-, and high-performing students (using teachers' reports of students' abilities) to think alouds and interviews.

After completing the curriculum that supported students with written scaffolds in developing evidence-based explanations about core ecological ideas, students were given a test containing 20 items. The test had items that gathered information about core ecological ideas, core ecological ideas fused with the practice of interpreting data, and core ecological ideas fused with the practice

This table shows schoolyard animal data collected using Cyber Tracker. Use the table to help you answer question 4.

School Yard Animal Data

Animal Name	Zone A	Zone B	Zone C
Pillbugs	1	3	4
Ants	4	6	10
Robins	0	2	0
Squirrels	0	2	2
Pigeons	1	1	0

Write a scientific explanation for the following question.

4. **Scientific Question:** Which zone likely contains the most habitats?

Make a CLAIM
Write a sentence that answers the scientific question.

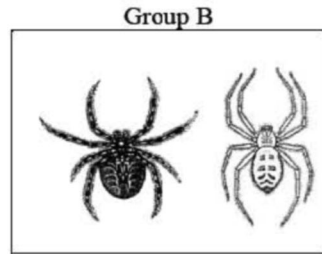
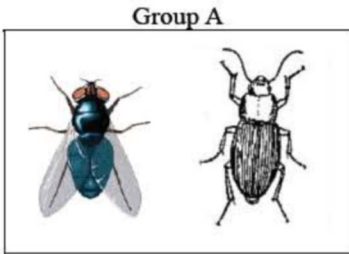
Give your EVIDENCE
Look at your data and find two pieces of evidence that help answer the scientific question.

- 1.
- 2.

Give your REASONING
Write the scientific concept of definition that you thought about to make your claim.

Figure 3. An Intermediate assessment item with practice scaffolds.

8. Shan and Nikki collected four animals from their schoolyard. They divided the animals into two groups based on the physical characteristics of the animals:



The next day, Shan and Nikki collected another animal from their schoolyard. It looks like this:



Shan and Nikki need to decide in which group this new animal belongs.

Write a scientific explanation for the following question.

Scientific Question: Does this animal belong in Group A or Group B?

Figure 4. A complex explanation item without scaffolds.

of scientific explanations. This analysis will focus on the eight open-ended items that fused core ecological ideas and scientific explanations. The interviews had two parts, a think-aloud section and a cognitive interview section. Common think aloud procedures were used in order to examine students' thought processes as they worked on the assessment tasks (Ericsson & Simon, 1993). After being instructed about the thinking aloud procedure, the interviewer modeled how to think aloud on one practice problem. Then the student practiced thinking aloud on a second practice problem. Following the practice, students thought aloud as they completed the assessment. The interviewer did not interact with the student as he or she completed the assessment except to remind the student to keep talking or to speak louder.

One limitation of think aloud data is that students will only say aloud what they have to think about when they interact with a task. For example, knowledge and skills that students know very well and/or are automatized will not necessarily be elicited by think-alouds (Chi, 1997; Ericsson & Simon, 1993). Thus, while think alouds provide a good look at how students interact with a

given task, they can only give us an imperfect picture of how students reason about certain situations. Therefore, after the student completed the assessment, the interviewer went back over the assessment with the student asking the student to clarify responses on items, to explain how they reasoned about an item, and/or and to talk about their perceptions of the items, for example, which they found difficult or easy and why. The combination of spontaneous think-aloud protocols and structured interview prompts allowed students to respond to items without intervention and at the same time allowed us to obtain information that was not volunteered in the unstructured think-aloud format (Kupermintz, Le, & Snow, 1999).

Coding and Analysis

In this article, we focus on students' abilities to fuse core ecological ideas with evidence-based explanations. We agree with Corcoran, Mosher, and Rogat (2009) that,

By treating the development of concepts and practices as analytically distinguishable, but intertwined, pathways. . . progressions can make this tension explicit and provide a basis for describing and assessing the empirically observable combinations of concepts and practices that actually show up in students' understanding and in their work. (p. 21)

Thus, we focus on examining the ways in which students utilize core ecological ideas as they develop explanations about specific ecological scenarios.

In order to analyze the think aloud and cognitive interviews, we followed standard procedures (DeBarger, Quellmalz, Fried, & Fujii, 2006; Ericsson & Simon, 1993). First the think alouds were transcribed. Following this, the transcripts were segmented first by assessment item and then segmented by idea units (DeBarger et al., 2006). These idea units are the smallest meaning phrases that contained identifiable information that could be used for analysis. The step of segmenting into idea units was driven by the constructs of interest (i.e., core disciplinary ideas, claims, evidence, and reasoning about ecological scenarios) and the codes that to be assigned (Chi, 1997; Ericsson & Simon, 1993). Each segment was then assigned one or more codes. For this study, we focus on the codes of core disciplinary ideas (based on the level of our core disciplinary ideas progression), claims, evidence, and reasoning. Table 2 provides examples of how we applied the think aloud codes. After coding, we examined the proportion of idea units coded for each construct within each item (Ayala, 2002; Yue, Ayala, & Shalveson, 2002). In addition, information from the cognitive interviews that followed the think-alouds were coded similarly to the think alouds and also examined to find patterns in which items students struggled with and which items students found easy, and reasons behind their responses.

Written assessment items were coded based on the three parts of the explanation—claim, evidence, and reasoning, with each aspect of the explanation needing to include correct core disciplinary ideas (which represents the fusion of these two progressions). Table 3A includes a generic rubric that is based on our practice progression. Table 3B provides the coding rubric for the assessment item in Figure 3. Claims and reasoning were coded dichotomously (correct or incorrect) and evidence was coded using a partial credit scale where students could receive credit for having correct evidence (2 points), correct but incomplete/insufficient evidence (1 point), or incorrect evidence (0 points). The rationale for this coding scheme comes from the generic practice-progression-based rubric that illustrates one level for including just a claim (Level 1); two levels for evidence [appropriate but insufficient (Level 2) and appropriate and sufficient (Level 3)]; and one level that includes reasoning (Level 4). At least two raters coded each item. Inter-rater reliability was established before coding at 90% agreement and checked after coding was finished to ensure that raters remained consistent throughout the coding process.

Table 2
Coding categories for think aloud and follow-up interview

Coding Categories	Definitions	Examples
Core ecological ideas		
Core disciplinary idea	For Example: C4: Organisms (animals) have different features. There are observable internal and external differences (some fly, some have scales, fur, wings, live in the water, etc.). Some of these differences are used to distinguish major groups.	“Insects have six legs and antennasso this bug is an insect with Group A”
Explanation		
Claim statement	Articulation of causal claim based on scientific question	“That bug should go in Group A”
Use of evidence	Use of data to support the claim	“Group A’s all got six legs and antennas”
Reasoning	Use of scientific principles to link the evidence to the claim	“Insects have six legs and antennas so this bug is an insect with Group A”

Note: Idea units could be given more than one code (e.g., content and reasoning).

We utilized an item response model that describes the relationship between students’ abilities and the probability of a certain response on an item. This analysis was used to provide information as to whether students interacted with the items in ways that we would have predicted based on the fusing of learning progressions. We first examined the dimensionality of the data. An exploratory factor analysis suggested that the data were unidimensional, indicating that we could proceed using a unidimensional model.³ A simple Rasch model (Rasch, 1960) includes one-person ability parameter and one item difficulty parameter in its formulation. Models within the Rasch family can be articulated using the Random Coefficients Multinomial Logit model (RCML) formulation (Adams & Wilson, 1996). This model can be represented as shown below where $\Pr(X_i = j)$ represents the probability of a response j to an item X_i .

$$\Pr(X_i = j) = \frac{\exp(b_{ij}\theta + a'_{ij}\xi)}{\sum_{k=1}^{K_i} \exp(b_{ik}\theta + a'_{ik}\xi)}$$

where $b_i = (b_{i1}, b_{i2}, \dots, b_{ik})$ is the scoring vector, $\xi = (\xi_1, \xi_s, \dots, \xi_n)$ is a vector of n free parameters, and a_{ik} is the linear combinations for $i = 1, \dots, I; k = 1, \dots, K_i$.

In this model, all item and student fit statistics fell between 0.75 and 1.25 [which according to Bond & Fox (2001) indicate good fits to the model]. In addition, an examination of the test characteristic curves for the model indicated our test provides good information for students with ability levels both below and above average (between -3 though $+3$ —a range of ability level into which almost all of our students fell). This means that the test provides adequate information about all students who took this test.

Table 3

A. Generic practice-progression-based rubric for evidence-based explanations

Level 4 Student constructs a complete evidence-based explanation
Level 3 Student makes a claim and backs it up with sufficient and appropriate evidence but does not use reasoning to tie the two together
Level 2 Student makes a claim and backs it up with appropriate but insufficient (partial) evidence
Level 1 Student makes a claim with either no evidence or with inappropriate evidence

B. Learning Progression-Based Coding Rubric: Fusing Core Ecological Ideas and Practice

Coding	Sample Student Responses
Claim Correct (1): The animal belongs in Group A Incorrect (0): The animal belongs in Group B	Claim: “Group A”; “A”; “The fly goes in group A”; “The bug goes in Group A”
Evidence Appropriate and sufficient Evidence (2 points): 2 or more pieces of evidence from below Partial (Insufficient) Evidence (1 point): 1 piece of evidence from below (can include additional inappropriate pieces of evidence) Possible evidence (based on the pictures): <ul style="list-style-type: none"> • The animal has 6 legs • The animal has 3 body parts • The animal has wings • The animal has antennae 	Partial Evidence: “it has 6 legs” “it has wings like the ones in A” “it has 3 body parts and it is a bug” “it does not have eight legs, it has six like in A” Appropriate and Sufficient Evidence: “The fly has 6 legs and wings like in A” “It has 6 legs and 3 body parts” “It has 6 legs, 3 body parts and wings”
Reasoning Includes Reasoning (1): E.g., <i>Explicit statement</i> that ties evidence to claim with a reasoning statement. I.e. “The animal and Group A are all insects and they share certain physical characteristics” No Reasoning (0): No explicit statement tying claim to evidence	Reasoning Statements: “It is an insect like the bugs in A and insects have 6 legs and 3 body parts” “The bug is an insect like Group A and they all have 6 legs and wings” “All insects have 6 legs, 3 body parts, and antennae”

Evidence to Support the Interpretation of Students' Responses

Support for a validity argument should include evidence that is focused on different parts of the interpretive argument (Kane, 2001). In this section, we present two types of empirical evidence. The findings from the think aloud analysis allow us to examine the extent to which our learning progression-based assessment tasks allowed students to respond with fused core disciplinary ideas-based explanations. The item response evidence allows us to determine the extent to which student responses to our assessment tasks can be mapped to knowledge that fuses information from our core disciplinary ideas and explanation progressions.

Think Aloud Evidence

To gain more insight into the factors that influenced how students constructed explanations about core ecological ideas, we examined think aloud and cognitive interviews with 20 students. We examined the proportion of idea units coded for the core disciplinary ideas associated with the item, claim, evidence and reasoning for each item. Figure 5 illustrates the findings.

Figure 5 illustrates that both the intermediate items (with explanation scaffolds) and complex items (open-ended items without scaffolds) provided students with opportunities to give evidence of their abilities with respect to fusing core disciplinary ideas with evidence-based explanations. In both of these item types, students utilized core disciplinary ideas consistent with that expected given the design of the item based on the learning progression. Students also provided claims, evidence, and (in smaller proportion) reasoning in the items, indicating that these items were eliciting the types of fused knowledge and evidence-based explanation for which they were designed. In most items, the proportion of codes for core disciplinary ideas was similar, but slightly higher, than reasoning.

In addition, we examined student responses based on their achievement level (4 high achieving students, 10 middle achieving students, and 6 low achieving students). Figure 6 illustrates these results.

In think alouds for both complex items and intermediate items the lowest proportion of idea units were coded for reasoning. This was especially true for the lower achieving students who rarely used reasoning in responding to the tasks during the think alouds (although there slightly more instances of reasoning in the intermediate items—when the scaffolds were present). The follow-up interviews allowed students to expand on their understandings about the items. In the follow-up interviews, the low achieving students continue to not provide much scientific reasoning, however, the middle and high achieving students verbalized much more about reasoning in the follow-up interview situations. This provides support for our explanation progression and associated rubric with reasoning being the most difficult aspect of developing evidence-based explanations.

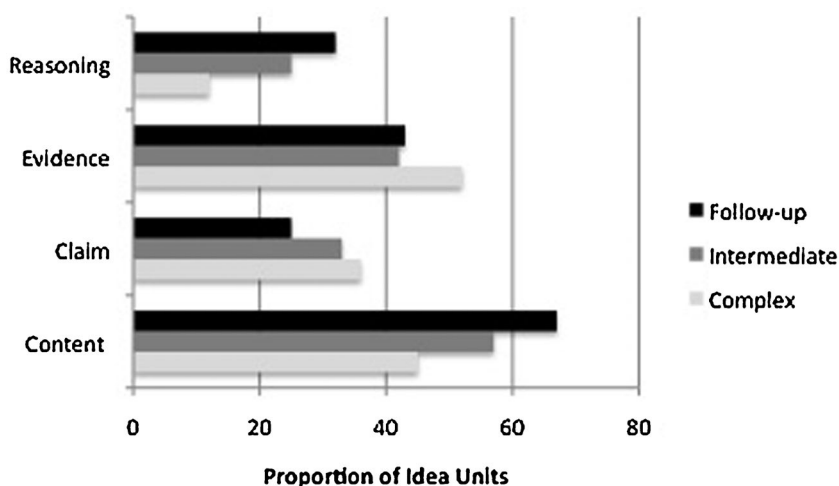


Figure 5. Think aloud and follow-up interview findings by item type.

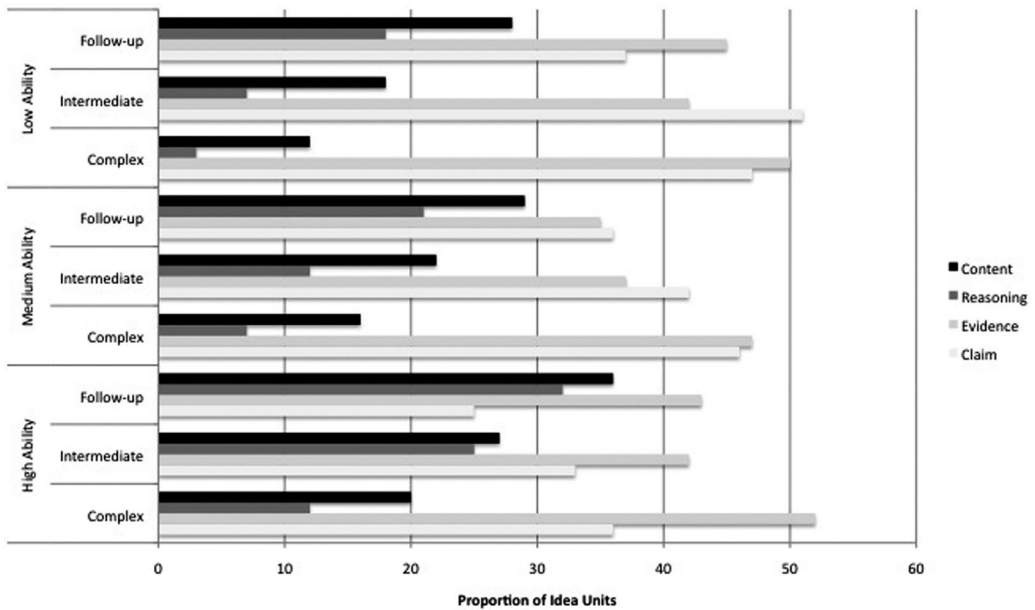


Figure 6. Think aloud and follow-up interview findings by item type and student ability level.

In order to better understand the patterns of students’ think alouds, we can examine the responses of a higher achieving student and a lower achieving student to several assessment items. Tatiyana’s teacher classified her as a “medium to high level student” and she had an ability level from our IRT analysis of 0.26 (slightly above average; see next section for information on the IRT analysis). Charity’s teacher classified her as a “lower level student with a lot of potential” and she had an ability level from our IRT analysis of -0.96 (below average). We can first examine the students’ responses to the intermediate item illustrated in Figure 3. A high level response to this item would include the correct zone (Zone B), with evidence that this zone has a higher richness (number of different types of animals) than the other school yard and reasoning that links the definition of habitats with why different types of animals need different habitats. Tatiyana responded to this item by saying,

Transcript and coding of Tatiyana’s think aloud about an intermediate item

Transcript Segment (Idea Unit)	Code
Um, one, three, four	(Reading the table)
So I think, I think its animal’s abundance um richness.	(Reading the table)
Oh so we have to go to Zone B	Claim
the first one’s one and four and one.	(Reading the table)
	Evidence
So and then I think this is the most, Zone B, claim.	Claim
Because all of the rest have zeros, some zeros	Evidence
and you have zone numbers and this one’s going to have a lot	Evidence
So I picked Zone, Zone B	Claim
because, um, my evidence and um reasoning is that it has more types of animals than the other zones	Evidence

Core Idea B3

In her response, Tatiyana does a thorough reading and interpretation of the table and uses mostly accurate evidence to support her claim, but she does not link this evidence back to the claim and use a scientific principle (in this case, why more types of animals would need different types of habitats) as reasoning to show why the evidence supports her claim. Parallel to this finding is that only one idea unit in her response was coded for a core disciplinary idea (B3: use of the ideas of richness and abundance in examining the biodiversity of an area). When the interviewer went over her responses with her after she completed the think aloud, she was asked how she thought about this question,

Transcript and coding of Tatiyana's follow-up interview about an intermediate item

Transcript Segment (Idea Unit)	Code
Interviewer: What did you think of when answering this question? Tatiyana: I used the graph to get my answer that B has the highest habitats.	Claim Evidence
Interviewer: So what do you think of when you think of habitats? What are habitats? Tatiyana: Like where animals live and drink and eat	Reasoning Core Idea B1
Interviewer: So why did you choose Zone B Tatiyana: 'cause like more um animals live there and so they need stuff to live and drink and eat and get it where they live, so B	Reasoning Core Idea B2 Core Idea B3 Core Idea B7

So, when asked probing questions by the interviewer, Tatiyana knew the reasoning for choosing Zone B, however, she did not include it when she was responding to the item even when the item prompted her to provide reasoning. She utilizes the idea of what a habitat is (core disciplinary idea B1) and why we can use measures of richness and abundance to predict the types of habitats in a given area (core disciplinary ideas B3, B4, and B7). The item was designed to gather information about core disciplinary idea B7, but the precursor ideas (B1, B3, and B4) all allowed her to use sufficient evidence and scientific reasoning in her evidence-based explanation. The responses that Tatiyana gave when prompted by the interviewer illustrate that she understood more than we would have expected given her written response and think aloud response. This item, while allowing Tatiyana some opportunities to demonstrate how she could fuse core disciplinary ideas into creating an evidence-based explanation, did not elicit her full understanding about using the core disciplinary ideas in creating an evidence-based explanation.

Charity's think aloud response to this intermediate item in the think alouds is as follows:

Transcript and coding of charity's think aloud about an intermediate item

Transcript Segment (Idea Unit)	Code
Make a claim. Um A, B, C. 1, 3, 4, 4, 6, 10	(Reading the table)
Make a claim. It is ants with the most.	Claim (incorrect)
Hmmm which zone likely contains the most	(Reading the question)
No wait, it is B with the most	Claim
Give your evidence. Because of the numbers in the data the school yard animal data	Evidence (getting at evidence)
Give your reasoning. Because of the numbers in the school yard animal data. Right.	(Using scaffolding) Evidence (getting at evidence)

In this think aloud, Charity reads through the responses in the table and originally comes up with the incorrect claim of "ants." However, when she goes back to read the question, she comes

up with the correct claim of Zone B. In providing evidence, she refers the “numbers in the table” without providing explicit evidence of richness or abundance. When she uses the scaffold to “give your reasoning,” she goes back to the table and uses the same idea that she used for her evidence. She does not explicitly use any core disciplinary ideas when responding to this item. When she was asked how she thought about the question, she responds with the following:

Transcript and coding of charity’s follow-up interview about an intermediate item

Transcript Segment (Idea Unit)	Code
Interviewer: What did you think of when answering this question?	
Charity: I used the numbers in the table	Evidence
Interviewer: So, how did you use the numbers in the table?	
Charity: I used the numbers in the table to get Zone B	Claim Evidence
Interviewer: Okay	
Charity: It says which zone likely contains the most habitats, so I used the table to get zone B	Claim Evidence
Interviewer: Okay, okay, so what did you think of when you were thinking about habitats?	
Charity: I was thinking about the numbers in the table	Evidence
Charity: and Zone B had the most numbers in the table for each animal so I chose B	Claim Evidence

In this follow-up interview, even with the prompts from the interviewer, Charity did not provide any scientific principles (i.e., core disciplinary ideas) as reasoning. In addition, she did not include any more evidence for how she used the numbers in the table as evidence for her claim. This indicates that Charity’s responses to the assessment item in the think aloud and the written assessment are good indicators of what she knows. In addition, it provides evidence for how lower level students may reason about this type of item: providing a claim, insufficient evidence, and no reasoning, which is consistent with our explanation progression and associated rubric.

Tatiana seemed to use the hints in the scaffolds to guide her when thinking aloud while responding to the tasks. For example, a task that used the same table as that presented in Figure 3, asked students which zone had the highest biodiversity. A high level response to this would include a claim that Zone B has the highest biodiversity; evidence that Zone B has the highest richness (number of different kinds of animals) and second highest abundance (total number of animals); and reasoning that richness and abundance both play a role in determining the biodiversity of a given area although, because biodiversity is a measure of the *variety* of organisms in an area, the richness variable in this item is more significant in determining total biodiversity. In the think aloud Tatiana states that:

Transcript and coding of Tatiana’s think aloud about a second intermediate item

Transcript Segment (Idea Unit)	Code
Zone B again	Claim
Evidence is the most kinds of animals, um richness	Evidence Partial reasoning Core Idea B3
Um two pieces of evidence, um both of richness and abundance—its got the highest richness and abundance, right?	Evidence (Using Scaffolding) Core Idea B5
wait its got the highest richness for sure, but abundance 9 plus um Zone C has higher abundance	Evidence
Well, Zone B still	Claim Core Idea B5
‘cause it has more animals than C.	Evidence
So I guess that’s better	Partial reasoning

In her response to the question, Tatiyana originally provides one piece of evidence, just the richness. However, then she appears to read the scaffold that prompts her to provide two pieces of evidence and then considers the abundance of animals. This finding was a trend in all of the think alouds that we conducted. Tatiyana uses appropriate core disciplinary ideas when coming up with her evidence. For example, she demonstrates understanding of what richness and abundance are through her use of data from the table. In addition, Tatiyana seems to begin to get at the reasoning in her response (i.e., “so I guess that’s better”), however, without elaborating on what she means by “better” the reasoning would not be considered adequate to support her claim. When asked after the think alouds, what she meant by “better,” Tatiyana said, “I don’t know . . . just better.” In this case, even when prompted to elaborate, Tatiyana was not able to demonstrate appropriate reasoning, again providing evidence that explicit scientific reasoning may be the most difficult part of an explanation for students to formulate and that the item elicited appropriate understanding from Tatiyana.

Charity’s response to this item provided similar evidence as the first item that she responded to. She provided a claim and insufficient evidence, and no reasoning, despite the written scaffolds.

In the complex explanation item presented in Figure 4 a high level response to this item would include Group A, with evidence that listed the physical characteristics similar to the fly and Group A (e.g., wings, antennae, three body parts, six legs) and reasoning that includes information that all animals classified as insects share common physical characteristics. Tatiyana responds by saying,

Transcript and coding of Tatiyana’s think aloud about a complex item

Transcript Segment (Idea Unit)	Code
And, so the fly enters in Group A my claim then evidence are maybe because he have wings	Claim Evidence Core Idea C4
yeah the fly has wings like the other ones	Evidence Core Idea C4

In her think aloud, Tatiyana provided one piece of evidence using the core disciplinary idea from C4, which was accurate, but not a sufficient amount, in order to support her claim. She was not able to provide accurate reasoning about why the physical characteristics of the fly linked it to Group A (instead she repeats evidence that use used before). When going over her responses with her after she completed the interviews, she was asked how she thought about this question,

Transcript and coding of Tatiyana’s follow-up interview about a complex item

Transcript Segment (Idea Unit)	Code
Interviewer: What did you think of when answering this question? Tatiyana: I looked at the pictures and the fly looks like the insects in Group A	Evidence Claim
Interviewer: So what do mean by insects in, uh, Group A? Tatiyana: Insects have no backbone and have six um legs and antennas and stuff like um the fly,	Evidence Core Idea C4 and C7
so, um, that’s why he goes in there with them.	Reasoning
Interviewer: You mean Group A and the fly all have these 6 legs and antennas Tatiyana: Yeah, you just look at the picture and um you can see it, they are all insects	Reasoning Core Idea C4 and C7

Tatiana, who did not provide sufficient evidence or reasoning in her initial response to the task, when prompted by the interviewer, displayed knowledge of creating a more complete evidence-based explanation. She used information about classification (core disciplinary ideas from C4 and C7) and provided appropriate and sufficient evidence about the physical characteristics of the fly that would place it in Group A and accurate reasoning that linked the fly to the other insects in Group A. This illustrates that this assessment task may not have allowed Tatiana to fully demonstrate all that she knew and could do.

Charity’s response to this item was very similar to Tatiana’s, she states: “This bug goes in group A. Umhhh it has wings.” She gives a claim and provides one piece of correct evidence in her response. In the follow-up interview, she states:

Transcript and coding of charity’s follow-up interview about a complex item

Transcript Segment (Idea Unit)	Code
Interviewer: What did you think of when answering this question? Charity: I looked at the bugs and saw that it goes in Group A. Interviewer: Okay, so why did you choose Group A? Charity: Because it looks like the other bugs.	Claim Evidence Core Idea C4
Charity: It can fly and it has wings	Evidence Core Idea C4
Interviewer: Okay. So it has wings and looks like the other bugs? Charity: Yes. The wings and bugs	Evidence Core Idea C4
Interviewer: Okay. Anything else? Charity: Nope.	

In this interview, she provides the same piece of evidence to support her claim that the animal belongs in Group A. However, even when asked by the interview if she wants to provide any more support (i.e., “Anything else?”) she does not include more evidence or provide any reasoning. Charity’s responses to the assessment item in the think aloud and the written assessment are good indicators of what she knows and provides evidence supporting our learning progression in that lower level students may respond by providing a claim, insufficient evidence, and no reasoning.

Difficulty Evidence

In order to better understand the extent to which students’ responses to items corresponded with knowledge that fuses information from our core disciplinary ideas and explanation progressions, we examined the characteristics of the eight constructed response explanation items and how students responded to these items. To do this, we used the difficulty parameters (b) of the items from the item response model. The difficulty continuum is set up as a logit scale and we set the mean of the item difficulty parameter to be 0. Items with negative difficulty are easier than average and items with a positive difficulty are more difficult than average. Table 4 illustrates that, on average, the claim is the easiest part of the explanation, easier than both evidence ($p < 0.05$) and reasoning ($p < 0.001$). Utilizing evidence to back up the claim is more difficult than the claim, but less difficult than providing reasoning ($p < 0.001$), and providing reasoning or scientific principles to link the evidence to the claim was the most difficult of all aspects of the explanation.

While there are clear patterns in Table 4 illustrating the difficulty of the different parts of scientific explanations, there is also a wide range of difficulty for each component of explanations. One possible reason for the wide range of difficulties could be linked to the scaffolding provided in the items. In particular there were three intermediate items that provided scaffolding hints (see Figure 3), while the other five complex items had no scaffolding. Table 5 provides the difficulty

Table 4

Difficult of explanation components (N = 8 explanation questions; 312 students)

	Average Difficulty	Range
Claim	-1.027	-3.07 to 0.612
Evidence	-0.545	-1.55 to 0.477
Reasoning	1.796	0.339 to 3.15

parameters of the items by scaffolding condition. This analysis shows that the average difficulty level for the scaffolded claims tended to be slightly higher than the unscaffolded claims, however, the difference is not statistically significant. Similarly, providing reasoning, regardless of scaffolding condition of the item, tended to be very difficult for students. However, in the difficulty parameters for providing evidence there is a large difference between items that provided scaffolding and those that did not provide scaffolding. Providing evidence in items that had scaffolding was much less difficult than in items without the scaffolding ($p < 0.01$). While the range of difficulty parameters in all items likely has much to do with the specific ecological core disciplinary ideas required to complete the task, given the closeness of difficulty parameters in claims and reasoning for scaffolded and unscaffolded items, the difference in average difficulty parameters for providing evidence between scaffolded and unscaffolded items stands out.

While the tables of averages for the different components of explanations provides an overview of the patterns that we saw, they do not give a full picture of the difficulty of each component of the explanation for each item. Another way of viewing the difficulty of items is to use a Wright Map that places students' proficiency levels on the same scale as item difficulties (Wilson, 2005). We used a Wright Map to illustrate the difficulty of the components of explanation items were relative to each other and to the students who took our test (Figure 7).⁴

The Wright map illustrates that the different components of items (claim, evidence, and reasoning about the different core disciplinary ideas) had a wide range of difficulties and was aligned with the range of ability levels of the students ("x"). This shows that, providing students with a wide range of ability levels good opportunities to respond to the items. The Wright map illustrates that for all items (except item 8),⁵ the claim was the easiest component and reasoning was the most difficult. In fact, we can see that few students used reasoning (i.e., an appropriate scientific principle or core disciplinary idea) to justify why the data counted as evidence in support of their claim [see the low number of students (x) at the upper end of the student continuum]. In addition the Wright map shows that scaffolding played the most role in supporting students in providing evidence (note that bold "ev" items are easier than non-bolded items), but that

Table 5

Difficulty of explanation components with scaffolding condition (N = 3 scaffolded items, 5 unscaffolded items; 312 students)

	Average Difficulty	Range
Claim (scaffolded)	-0.96	-1.97 to 0.50
Claim (unscaffolded)	-1.05	-3.07 to 0.612
Evidence (scaffolded)	-1.475	-1.55 to -1.40
Evidence (unscaffolded)	-0.23	-0.997 to 0.477
Reasoning (scaffolded)	1.735	0.34 to 3.15
Reasoning (unscaffolded)	1.78	0.339 to 2.788

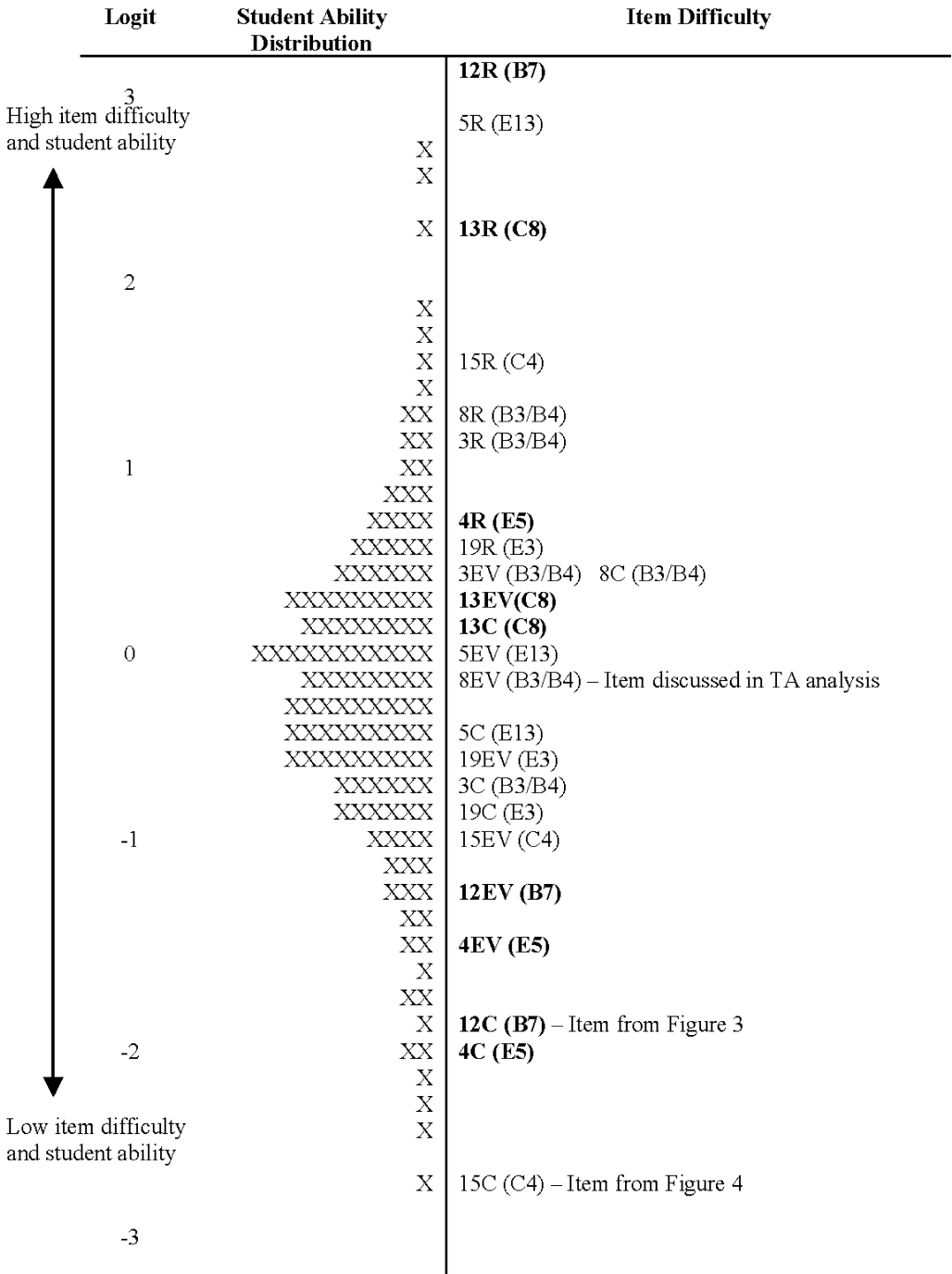


Figure 7. Wright map. Note. Each x represents 2.4 cases. C, claim, EV, evidence, R, reasoning. Content information in parentheses. Bold items contain scaffolding.

scaffolding did not seem to play a role in the difficulty of the reasoning component of explanations (there is no pattern in difficulty level between items with scaffolding and those without scaffolding). This is consistent with the patterns reported above and it illustrates a consistency between items for the ordering of difficulty of the components of explanation.

The pattern in core disciplinary ideas of the items is a bit messier to discern. In the think alouds, the most double coded idea units were those that contained both core disciplinary ideas and reasoning. This makes sense given that reasoning is the aspect of the explanation that uses a scientific principle or core disciplinary idea to tie the evidence to the claim. Thus, we can examine the pattern in the difficulty parameter of the reasoning aspect of the items to get a better sense of the patterns in core disciplinary ideas. The pattern that emerges is that within each strand of the core disciplinary ideas progression (classification, ecology, and biodiversity) the higher levels of core disciplinary ideas tended to be more difficult than the lower levels (e.g., B7 was more difficult than B3 or B4 and C8 was more difficult than C4). This is what we would have expected given our learning progression. However, the pattern does not hold for all aspects of the explanation. For example, the claim with the lowest difficulty was for C8 while the claim for C4 had a much higher difficulty level.

Evaluation of the Validity Evidence

Below we explore what we learned about how well our items elicited student responses consistent with knowledge that fuses information from the core disciplinary ideas and explanation progressions.

Core Disciplinary Ideas

Findings from the think alouds illustrate that each item elicited core disciplinary ideas consistent with what would be expected given the progression on which they were based. In addition, the think alouds showed that, at times, students also used core disciplinary ideas from the progression that were precursors to the core disciplinary ideas in the item. For example, Tatiyana used core disciplinary ideas from B1, B2, B3, and B7 in an item that was designed to gather information about B7. This shows that the item allowed for students to show mastery of the current and precursor core disciplinary ideas when crafting an evidence-based explanation. Our goal is not to place students at a given level on our core disciplinary ideas progression. Rather, we use the core disciplinary ideas progression as a means to sequence our curricular units and allow students appropriate opportunities to fuse core disciplinary ideas with their evidence-based explanations.

The findings about core disciplinary ideas from the difficulty analysis were more nuanced. When examining the reasoning component of the item, there was alignment within core disciplinary ideas strands (classification, ecology, and biodiversity) with higher-level ideas being more difficult. However, this pattern did not hold for claims or evidence. This illustrates the importance of examining how students fuse core disciplinary ideas into the different aspects of their explanations.

Claims

All of our items elicited claims from students (and many of these claims used correct core disciplinary ideas), matching our hypothesis that claims may be relatively easy for students to craft since they are, in this case, usually an answer to a scientific question that is posed to them. In both the difficulty analysis and in the think alouds and interviews, students were able provide a claim, and it was the component of the explanation that was most likely to be answered accurately. There was a range of empirical difficult parameters for the claim (see Figure 7). The main factors in whether students are able to create a correct claim are whether they understand what the

question is asking as well as their knowledge of the core disciplinary ideas implicit in the scientific situation (Gotwals et al., 2012). In addition, in situations that require students to interpret a table, graph, or some other inscription, the ability to interpret the data accurately will play a role in whether students are able to make an accurate claim (Gotwals & Songer, 2006). While some students may create claims that do not answer a scientific question or leave the claim blank since they are unsure of the correct response, the claim component of the explanations was the least problematic in terms of linking students' responses to our learning progression for evidence-based explanations.

Evidence

All of our items also provided students with the opportunity to use evidence in backing up their claim. Our hypothesis was that, after creating claims, the next most difficult component of creating explanations is to provide sufficient and appropriate evidence to support the claim. This was generally true based on our difficulty analysis and the think alouds. In the think alouds, the lower achieving students were less likely to include sufficient and appropriate evidence. In the difficulty analysis, there was a large range in the difficulty of providing evidence for claims, with middle and higher achieving students more likely to include appropriate and sufficient evidence. For example, Tatiyana provided more evidence in the think alouds and written examples than Charity.

Providing evidence may come naturally in some situations (e.g., often students will state a claim and then say "because..." and list evidence). However, determining what counts as appropriate evidence in the given scientific situation and how much evidence is sufficient to support a given claim may be more complicated for students to determine. McNeill (2011) found that many students think of evidence in terms of "support[ing] an answer to a question" (p. 811), but do not always view data as evidence. Providing scaffolding for students as to how much evidence and what counts as evidence in a given situation may guide students as they are learning how to create coherent explanations. Our findings indicate that the scaffolding in our assessments played a significant role in supporting students in providing more appropriate and sufficient evidence. Based on our difficulty analysis, providing appropriate and sufficient evidence for tasks that included scaffolding was significantly easier. In addition, when scaffolded in the follow-up interviews, we can see that the students tend to provide more appropriate and sufficient evidence. Thus, our hypothesis that evidence would be more difficult than claims and less difficult than reasoning was also supported with data. In addition, the scaffolding that we provided in the assessment prompts appears to provide significant support for students in creating strong evidence-based explanations that have appropriate and sufficient support, providing evidence for our design decision to use scaffolds in our assessment tasks.

Reasoning

While less likely to be written or verbalized, our items allowed all students to provide scientific reasoning to link their evidence to the claim. Our final hypothesis about our learning progression tasks was that providing explicit reasoning as to why the evidence counts as reasoning is likely a more difficult step in the process of learning to create scientific explanations. This hypothesis was supported by our difficulty analysis, with the reasoning component of the explanations being the most difficult for students. In addition, in the think alouds, verbalizations of reasoning using core disciplinary ideas occurred for the smallest proportion of time, especially for the lower achieving students.

Our hypothesis about our scaffolding supporting students in developing reasoning in their written explanations had mixed results. In the written items, scaffolding did not seem to influence

the difficulty parameter of items. In addition, we saw that in the think alouds, Tatiyana did not always provide reasoning, and even when she did read the prompt for reasoning, it did not always lead her to provide adequate reasoning to support her claim. In the follow-up interviews, however, when asked questions about how they were thinking about certain aspects of the item or their response, students were much more likely to provide reasoning and illustrate evidence that they had.

Evaluation of Validity Evidence

The evidence that we gathered indicates that our items provided students with opportunities to fuse core disciplinary ideas into evidence-based explanations about the given scientific scenarios. Higher ability students tended to score higher on the items than lower ability students (providing appropriate and sufficient evidence and scientific reasoning), showing that the items were able to distinguish between students at different levels of learning to craft fused core disciplinary ideas into evidence-based explanations.

Our findings also indicate that some students may understand how to support their claims through evidence and reasoning, but not make it explicit in their written (or verbal) explanations. Perhaps this is because they think that it is “a given” or a shared understanding between themselves and the audience to whom they are explaining. Tatiyana may not have thought that providing a definition of insects was necessary since the interviewer may have already known this. Similarly, Charity, may not have thought that it was important to include the evidence that all of the “bugs” had wings, since it was obvious to her. Alternatively, students may have some core disciplinary ideas that could be used help them provide evidence to support a claim or to provide reasoning in a given situation, but the core disciplinary ideas may not be strong enough to provide a clear link between their claim and what they believe is evidence to support that claim. For example, Tatiyana seemed to know that the richness data were somehow more important or “better” when crafting her explanation, but could not elaborate what she meant by this. We would have hypothesized that our scaffolding would support students in realizing that they need to make their reasoning explicit and to provide hints about what counts as reasoning in a given situation. However, this was not the case. Based on this and other data that we have collected (see Songer & Gotwals, 2012) we will work on developing better ways of supporting students in the reasoning component of their explanations, but also in all aspects of their progression towards creating sophisticated scientific explanations that fuse core disciplinary ideas with scientific practices.

Perhaps more challenging for our validity argument were the responses to the interview questions. At times, when the interviewer probed Tatiyana, Charity, and other students to elaborate on their responses, they provided evidence that they knew more core disciplinary ideas and more about constructing evidence-based explanations than came through in the written and think aloud responses. Tatiyana did not use the word “insect” in either her written work or in her think aloud interview about grouping invertebrates. However, when the interviewer asked her what she thought about when answering the question, Tatiyana demonstrated that she both knew what insects were (which is part of the reasoning) and that she could provide more evidence to support her claim (all of the characteristics of insects). These findings indicate that, perhaps, reasoning is not always as difficult as we hypothesized and that it might be the nature of our assessment tasks and think alouds that did not indicate to students, somehow, that they needed to include explicit links between the core disciplinary ideas that linked their evidence to their claim. Thus, this might indicate that our explanation progression should not include reasoning at the highest level, rather we might need to consider all of the types of support that students may provide in a given explanation and the difficulty of piecing together all of the components into a coherent whole.

Discussion

The Framework for Science Education Standards (NRC, 2011) stresses that “because R&D [research and development] on learning progressions in science is at an early stage, many aspects of the core ideas and their progressions over time . . . remain unexplored territory . . . [A]n especially important line of inquiry should involve learning progressions that embed the core ideas and practices spelled out in this document” (p. 315). However, the Framework also states that while assessment is a crucial component of instantiating the fusion of core ideas and practices (and cross-cutting concepts), “[D]etails about the design of assessments for any given purpose of context are beyond the scope of this framework” (p. 264). Having an assessment system that is developed using (a) empirical data of student learning, (b) tasks designed to elicit observations of fused core disciplinary ideas and practices from students, and (c) analysis designed to provide insights on students’ successes and difficulties in generating fused core ideas plus practices knowledge is essential if we are to learn more about the intricacies and complexities of the ways in which students develop complex reasoning in science (e.g., Pellegrino, 2012). This article provided an example of how we gathered validity evidence for the extent to which our assessment tasks map onto knowledge that fuses information from our core disciplinary ideas and explanation progressions.

Kane (1992) stated that “. . .one possible criterion for evaluating validation research is the extent to which the research improves both the interpretation (by making it clearer, more solidly based, and more accurate) and the test (by eliminating flaws and sources of error)” (p. 532). This study may qualify as a formative validity study for our research project to better understand the ways in which our assessment tasks allowed responses that could be mapped to the fusion of information from our core disciplinary ideas and explanation progressions. In this sense, our study provided evidence as to the strength of the interpretation that we can make about our assessment tasks soliciting evidence related to the knowledge that fuses information from our core disciplinary ideas and explanation progressions, thus allowing revision to our assessment tasks and coding rubrics for the future. In addition, the types of information that students provided also gave us insight into how to revise and improve our core disciplinary ideas and explanation progressions.

The Importance of Validity Arguments for Learning Progression-Based Assessment

There are many challenges associated with developing assessment tasks that can elicit student responses about learning progressions. The challenges make developing a validity argument about the items themselves particularly important. Much in the same way that we (in our research project and in the NGSS) are working with students to develop their abilities to create and support claims using evidence and reasoning, using an argument-based approach to assessment supports our work to develop claims about what our assessment can and cannot do and systematically collect evidence to support these claims. We found that our assessment tasks allowed students to draw on the core disciplinary ideas and the practice of developing evidence-based explanations that they were designed to do. Assessment tasks that take a “learning progressions stance” (Alonzo, 2012) need to allow students at multiple levels opportunities to illustrate what they know. Our scaffold-rich assessment tasks allowed lower and middle ability students opportunities to illustrate their ability to provide evidence to back up their claims in a manner that would not be possible with items that did not contain scaffolds. However, we found that the scaffolding that asked students to provide reasoning for the task did not have the desired effect; in fact, this scaffolding did not seem to make a difference in assisting students in generating valid reasoning statements. Our conclusion is that this piece of our scaffolding was ineffective and needs to be reconsidered. While we do not

want to add more reading into the assessment tasks, especially for our younger students (Songer & Gotwals, 2012), thinking about effective ways of eliciting fused core disciplinary ideas in reasoning statements is important. One suggestion might be to examine the ways in which effective teachers verbally probe students to provide reasoning for their explanations and attempt to translate this into written prompts or hints (e.g., Songer, Shah, & Fick, in press). This manner of refining our assessment task illustrates how we can formatively use the validity evidence to better our tasks, as Kane (1992) suggested.

The Importance of Validity Arguments for Learning Progressions

At this stage, learning progressions research may be characterized as an “epistemic enterprise” (L. Schauble, personal communication, July 26, 2011) in that the field is both producing and refining knowledge about learning progressions at the same time. Research to gather evidence about the validity of learning progressions requires that assessments are designed to reliably capture the nature of students’ understandings as well as matching this knowledge to levels of a learning progression. However, gathering validity evidence using learning progression-based assessments can pose some difficulties if students’ responses to the assessments do not align with what is expected based on the learning progression. Much of this early work in using learning progressions to inform assessment (or curriculum and other uses) must therefore involve simultaneously gathering evidence about both the validity of learning progression-based assessment items and the match of student responses to the learning progression templates. Conflicts may arise, however, when simultaneously examining evidence about both the construct (i.e., the knowledge in the learning progression) and the items used to assess the knowledge in the learning progression. If, for example, students’ responses to an item do not match with what would be expected for that knowledge at that location of the learning progression, it is hard to know whether this information provides evidence that the item was not a “good” item that provides valid information on the construct or whether the outcome information provides evidence that the knowledge in the learning progression does not accurately capture students’ capabilities relative to the construct. These questions arise whether we have evidence to help us determine whether the knowledge in the learning progression is an appropriate construct or the assessments designed to tap into the knowledge represented in the learning progression need improvement. Thus, a validity argument must be carefully evaluated so that any interpretive decisions about what the data show can be supported.

In this study, the think aloud data showed that students drew on core disciplinary ideas in reasoning about the ecological scenarios, illustrating that our core disciplinary ideas progression has allowed students to build on the basic knowledge and use this knowledge in considering more complex scenarios. In addition, the difficulty findings illustrate that generating claims about core disciplinary ideas do tend to be the easiest component of explanations for students to generate, with generating evidence as the next most difficult and generating reasoning as the most difficult aspect of the explanation for students to create. These results are consistent with others’ findings (e.g., McNeill et al., 2006) and they provide information that the path of our practice progression is consistent with students’ written responses. However, our data did not suggest clear patterns in the difficulty of the core disciplinary ideas when fused to explanation building. Such an outcome suggests that the work of constructing fused knowledge by students is complicated by both the difficulty of the core disciplinary idea and the amount of prior knowledge and support associated with the student and each assessment task. More recently, our project has developed a single learning progression that represents knowledge that is a fusion of core disciplinary ideas and practices (as referenced in Duschl, Maeng, & Sezen, 2011) to attempt to gather more empirical

information on the developmental trajectory associated with the development of fused knowledge over time and topic.

Our results illustrate the importance of gathering multiple sources of data about what students know and can do when gathering evidence about a learning progression and learning progression-based assessment items. If we were to only rely on the results from our difficulty analysis, we would not have examined the nuance in students' abilities to use reasoning and evidence. Thus, especially in these times of the early work on learning progressions, it is imperative that we gather rich and varied sources of data about the nuances of students' understanding and learning through written work, think-alouds, interviews, and other data sources (such as curricular interventions).

Overall, our results suggest that learning progressions can serve as useful templates for the development and analysis of assessment items that can generate information about the ways students both succeed and struggle in developing scientific knowledge that fuses core disciplinary ideas with science practices. However, while learning progressions (such as ours) are often displayed as hierarchical levels, most researchers agree that learning does not customarily follow linear, sequential steps of development (Songer & Gotwals, 2012) and that students need to revisit ideas through repeated guidance, reflection and multiple exposures in order to develop the sophisticated scientific knowledge that fuses core disciplinary ideas with practices (Songer et al., 2009). Thus, the construct of the learning progression is an idealized sequence rather than a stepwise path.

Learning progressions can be used to link curriculum, assessment, and professional development in order to provide students and teachers with a coherent experience. However, we postulate that while it is possible to gather validity evidence about the material represented in a learning progression (e.g., core disciplinary ideas and science practices), we cannot make a claim that our learning progression is the only (or the best) progression that represents how students learn to craft evidence-based explanations about core ecological core disciplinary ideas. Rather, we suggest that validity studies are an important component of the work necessary to gather empirical evidence on the challenge associated with fostering and assessing students' fused knowledge development.

Notes

¹In a recent issue of *Measurement*, Newton (2012) proposes a clarification of the construct of validity, with a more dichotomous decision of valid or not valid. However, the comments following his argument push back on this binary notion of validity (e.g., Mislevy, 2012).

²The assessment tasks that were designed for this study were not based on a single fused learning progression. Rather, the data for this article are based on a construct that fuses the core disciplinary ideas from our three strands (i.e., classification, ecology, and biodiversity) with the scientific practice of evidence-based explanations at key junctures throughout the learning process to create performance expectations, which in turn, inform the design of curricular and assessment tasks.

³Although we used a unidimensional model, this does not preclude us from thinking about multiple aspects of students' understandings (e.g., core disciplinary ideas AND practices) influencing their responses. In fact, "any of these IRT models we use are great oversimplifications of what is happening cognitively. There are many more aspects of knowledge and skill that examinees are bringing to bear in the tasks. All the IRT model is doing is looking for major, joint, relationships of patterns across items where responses have co-occurrences that can be modeled by some number of dimensions along which examinees might be characterized in terms of. If examinees tend to get more or fewer items right and we can use a single variable to approximate

the response matrix adequately, then. . . a unidimensional model fits—no matter how much is going on cognitively. . .” (R. Mislevy, personal communication, January 9, 2013).

⁴Evidence was scored polytomously, but we averaged the step difficulties to provide one difficulty parameter for evidence for each item.

⁵In item 8, students used the chart from Figure 3, to answer, “What zone had the highest biodiversity?” Many students chose Zone C because it had the highest number of animals (an incorrect claim), not considering the richness of animals (which would have led them to Zone B). Thus, these students were able to choose the wrong zone (Zone C), but still get credit for partial evidence (examining the abundance of animals).

References

- Adams, R. J., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. Wilson (Eds.), *Objective measurement III: Theory into practice* (pp. 143–166). Norwood, NJ: Ablex.
- AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Alonzo, A. C. (2012). Eliciting student responses relative to a learning progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 241–254). Rotterdam, The Netherlands: Sense Publishing.
- Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, 93, 389–421.
- American Association for the Advancement of Science [AAAS]. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.
- Anderson, C. W., Alonzo, A. C., Smith, C., & Wilson, M. (2007, August). *NAEP Pilot Learning Progression Framework*. Report to the National Assessment Governing Board.
- Ayala, C. C. (2002). *On The Cognitive Validity Of Science Performance Assessments Using A Research Based Knowledge Framework*. Unpublished Doctoral Dissertation, Stanford University, Palo Alto, CA.
- Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education*, 94, 765–793.
- Berland, L. K., & Reiser, B. J. (2010). Classroom communities adaptations of the practice of scientific argumentation. *Science Education*, 95, 191–216.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in human sciences*. Mahwah, NJ: Erlbaum.
- Bricker, L., & Bell, P. (April, 2007). *Um . . . since I argue for fun, I don't remember what I argue about: Using children's argumentation across social contexts to inform science instruction*. Paper presented at the National Association of Research in Science Teaching, New Orleans, LA.
- Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, 6(3), 271–315.
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2), 121–152.
- College Board. (2009). *Science College Board Standards for College Success*. Available: <http://professionals.collegeboard.com/profdownload/cbscs-sciencestandards-2009.pdf> [June 2011].
- Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report #RR-63). Philadelphia, PA: Consortium for Policy Research in Education.
- DeBarger, A. H., Quellmalz, E., Fried, R., & Fujii, R., (2006). *Examining the validities of science inquiry assessments with cognitive analyses*. Paper presented at the Annual meeting of the American Educational Research Association, San Francisco, CA.
- Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progression and teaching sequences: A review and analysis. *Studies in Science Education*, 47(2), 123–182.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Gotwals, A. W., & Songer, N. B. (2006). Measuring students' scientific content and inquiry reasoning. In S. A. Barab, K. E. Hay, & D. T. Hickey (Eds.), *The Proceedings of the Seventh International Conference of the Learning Sciences* (pp. 196–202). Mahwah, NJ: Lawrence Erlbaum Assoc.

Gotwals, A. W., & Songer, N. B. (2010). Reasoning up and down a food chain: using an assessment framework to investigate students' middle knowledge. *Science Education*, 94, 259–281.

Gotwals, A. W., Songer, N. B., & Bullard, L. (2012). Assessing students' progressing abilities to construct scientific explanations. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 183–210). The Netherlands: Sense Publishing.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342.

Kupermintz, H., Le, V.-N., & Snow, R. E. (1999). *Construct validation of mathematics achievement: Evidence from interview procedures*. Los Angeles, CA: Stanford University: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Liu, O. L., Lee, H.-S., Hofstetter, C., & Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures and evidence. *Educational Assessment*, 13, 33–55.

McNeill, K. (2011). Elementary students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year. *Journal of Research in Science Teaching*, 48(7), 793–823.

McNeill, K., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 233–265). New York: Taylor & Francis.

McNeill, K., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences*, 15(2), 153–191.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.

Mislevy, R. J. (2012). The case for informal argument. *Measurement*, 10, 93–96.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.

National Assessment Governing Board. (2008). *Science framework for the 2009 National Assessment of Educational Progress*. Retrieved from National Assessment Governing Board website: <http://www.nagb.org/publications/frameworks/science-09.pdf>

National Research Council. (2001). *Knowing what students know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.

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

National Research Council. (2011). *A Framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement*, 10, 1–29.

Pellegrino, J. W. (2012). Comment: Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching*, 49(6), 831–841.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Ruiz-Primo, M. A., Li, M., Tsai, S.-P., & Schneider, J. (2010). Testing one premise of scientific literacy in classrooms: Examining students' scientific explanations and student learning. *Journal of Research in Science Teaching*, 47(5), 583–608.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39(5), 369–393.

Songer, N. B., & Gotwals, A. W. (2012). Guiding explanation construction by children at the entry points of learning progressions. *Journal for Research in Science Teaching*, 49, 141–165.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning in biodiversity. *Journal of Research in Science Teaching*, 46(6), 610–631.

Songer, N. B., Shah, A. M. & Fick, S. (in press). Characterizing teachers' verbal scaffolds to guide elementary students' creation of scientific explanations. *School Science and Mathematics*.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching*, 46(6), 699–715.

White, B., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, 16(1), 3–118.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, New Jersey: Laurence Erlbaum Associates, Publishers.

Yue, Y., Ayala, C. C., & Shavelson, R. J. (2002). *Student's problem solving strategies in performance assessments: Hands on minds on*. Paper presented at the Paper presented at the American Educational Research Association, New Orleans, LA.