Measuring students' understanding of science and engineering practices, disciplinary core ideas and crosscutting concepts through an embedded assessment system

Anna MacPherson[1], Dora Kastel[1], John Howarth[2], Barbara Nagle[2], Maia Wilcox[2], & James Short[1]

The American Museum of Natural History [1]
The Lawrence Hall of Science [2]

This material is based upon work funded by the National Science Foundation under Grant # NSF DRL 1418235. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Contact info:
Anna Macpherson
The American Museum of Natural History
79th Street and Central Park West, New York, NY 10024
Phone: (212) 769-5261
amacpherson@amnh.org

Introduction

The Next Generation Science Standards (NGSS, Achieve, 2013) require a re-conceptualization of science teaching. A focus on *assessments* may be one way in which to shift science teaching toward practices that align with the new Framework for K-12 Science Education (NRC, 2012) and the NGSS (Achieve, 2013). Teachers use assessments to read the values of the science curriculum (Sunal & Wright, 2006) and they identify assessments as an important influence on how they choose the content to teach (Au, 2007; Hannaway & Hamilton, 2008; Weiss, Pasley, Smith, Banilower & Heck, 2003). Scholars in the fields of science education and educational policy agree that implementing better assessments, specifically ones that focus on problem-solving and critical thinking, could improve the quality of science education. Shavelson and Baxter (1991) argue that changing the tests may have a direct impact on how science is taught. According to them, "Teaching to tests that reward diverse problem solving . . . may directly influence teachers' instructional decisions" (p. 81)**.**

However, being able to use assessments as levers to improve instruction requires the *existence* of such assessments; specifically, assessments that measure the types of complex competencies we want to develop in students. Assessments are needed that measure a combination of students' understanding of disciplinary core ideas, eight scientific practices and seven cross-cutting concepts. This will likely require multi-component tasks that tap into all three dimensions, efficient methods for delivering, scoring and interpreting the results of the items, and studies to determine their validity and reliability (NRC, 2014). This is a tall order for the field.

This paper describes efforts to conceptualize, build, and study an assessment system embedded in a newly developed NGSS-aligned curriculum unit for middle school students. The unit, titled *Disruptions in Ecosystems: Ecosystem Interactions, Energy, and Dynamics* (the "*Disruptions in Ecosystems* unit"), addresses a bundle of performance expectations through a series of instructional sequences based on the BSCS 5E instructional model. The unit aligns with the NGSS performance expectations in Life Science Core Idea 2: Interactions, Energy, and Dynamics Relationships in Ecosystems and is focused on helping students answer the question, "How does a system of living and nonliving things operate to meet the needs of the organisms in an ecosystem?" The curriculum unit addresses all three sub-ideas: Interdependent Relationships in Ecosystems; Cycles of Matter and Energy Transfer in Ecosystems; and Ecosystem Dynamics, Functioning and Resilience. The curriculum supports the content primarily through the science and engineering practices developing and using models, constructing explanations and designing solutions and engaging in argument from evidence. The crosscutting concepts of patterns, stability and change, energy and matter, and cause and effect are emphasized in various chapters.

What is "responsible" assessment?

A major strand of research and development work on this project involves the development of teacher and student measures of learning. Throughout the process of drafting, piloting, refining, and testing these measures, we frequently return to the idea that we are seeking to design *responsible assessments*. We are in an age of accountability in education, and assessments occupy an outsized role in the system. Thus, we are striving in our work to create responsible assessments, which we define as meeting three criteria: (1) they are valid, reliable and fair (AERA & APA, 2014; (2) embedded in the work of learning and teaching; and (3) accessible to diverse students. What do these three components mean generally, and what do they mean in the context of this work?

Reliable, Valid, Fair

Validity is the evaluative judgment of the degree to which theory and empirical evidence support the adequacy and appropriateness of interpretations and consequences of a performance (score) on a test (Messick, 1989). In other words, does a test measure what it purports to measure and are the scores from such a test used for justified ends? What does that mean for us doing this work? Given that our student assessments are supposed to measure three-dimensional science learning, including disciplinary core ideas, science and engineering practices and crosscutting concepts, our assessments should not consist of multiple choice items that test basic recall only. Furthermore, given that NGSS-aligned assessments should reflect a developmental perspective (the Framework outlines a developmental progression for core ideas and practices, (NRC, 2012)) and the score on any one item, any one assessment, or even the suite of assessments, should not be used as a high-stakes measure. Rather, scores should be used to (1) provide feedback to student and teachers on progress toward goals along a learning progression (2) guide future instruction and (3) be used as one of many pieces of evidence to support an argument about students' proficiency in science.

Test *reliability* refers to the degree to which an assessment is consistent and stable in measuring what it is intended to measure. An assessment is reliable if it is consistent within itself and across time (AERA & APA, 2014). In this study, there is unavoidable tension between maximizing validity and reliability. Assessments of student learning with high validity will likely be multi-component performance tasks, which have notoriously low test-retest reliability and inter-rater reliability. Thus, this research may provide insight into the best ways of investigating reliability (e.g. which measurement models provide the most helpful estimate) and possibly how to maximize reliability while maintaining high validity and keeping costs (time and resources) manageable.

Test fairness includes the idea that a test should not be biased against certain test-takers, but also that the scores of any one test should not disadvantage any particular group. That is, scores or results from assessments should be *used for good*. In the case of this study, "used for good" means that results from the assessments should be used to improve teaching and learning, and not for placement, promotion, or high-stakes judgments of teaching

Embedded in the work of learning and teaching

Why is *embedded in the work of learning and teaching* important for our definition of responsible assessment? Simply put, assessments drive instruction, and so if our assessments are not deeply embedded in learning and teaching, then the tests themselves will shift instruction away from what matters. Research shows that the *structure* and *content* of assessments have a powerful influence on the way that science teaching and learning occur in classrooms in the United States (Hamilton & Hamilton, and 2008). Accountability policies lead teachers to focus on particular subject areas and particular types of instructional practices. Moreover, teachers focus on skills specific to assessment and testing procedures. In such a context, assessments become important signifiers of the intent of curriculum. As Wiliam (2010) argues, assessments "operationalize" constructs — the structure of assessments communicates to teachers what the desired practice looks like, and accompanying scoring guides provide insight into the types of performances that are expected from students.

So, our argument is that if we build a high quality assessment (i.e. valid, see argument above), and that assessment is embedded in teaching and learning, then "teaching to the test" becomes a good thing. It is not reasonable to tell teachers "don't prep for this test," and in fact we may be actually losing out on an important opportunity to help students and teachers understand NGSS by telling them not to pay attention to the tests. We want to design responsible assessments that are embedded in instruction, such that attending to the demands of the test help students and teachers understand what is expected. Our ideal assessment is one for which "teaching to the test" doesn't mean a narrowing of curriculum, fragmentation of knowledge and a shift-toward teacher-centered pedagogy (Au, 2007).

Accessible to diverse students

Finally, we believe that responsible assessments are accessible to diverse students. We aim to reduce test bias, though this is a highly specific psychometric property that we will not be able to investigate until the final year of the project when we have a large, diverse sample. However, we may gain insight into sources of bias during piloting and cognitive interviews. More generally, the assessments developed for this project should be able to be adapted to accommodate a variety of language and learning needs. Why is "accessible to diverse students" such an important part of our "responsible assessments" argument? If NGSS are going to be successful nationally, the assessments developed to measure proficiency must be accessible to all learners and should not be biased against particular groups.

Given the need for assessments of learning for the NGSS, and our focus on responsible assessment, this paper addresses the following specific research question:

- How can we develop responsible assessments of middle school students' learning during an NGSS-aligned curriculum about ecosystems?

Methods

Development of assessments

Educators at the Lawrence Hall of Science and the American Museum of Natural History collaboratively designed the *Disruptions in Ecosystems* curriculum unit and professional development using the *Five Tools and Processes for NGSS* (http://www.amnh.org/explore/curriculum-collections/five-tools-and-processes-for-ngss , AMNH, BSCS, WestEd, 2015). The *Five Tools and Processes* were developed using the principles of backward design (Wiggins & McTigue, 2005) and assessment-centered teaching (DiRanna, et al., 2008). In short, the *Five Tools and Processes* are a systematic process for professional development leaders to work with teachers to create NGSS-aligned curriculum, instruction and assessments. By using the *Five Tools and Processes,* teachers can translate science concepts, practices and performance expectations (NRC, 2012; Achieve, 2013) into multiple instructional sequences that form an NGSS unit. This is followed by a more in depth plan for one instructional sequence and assessment task to provide evidence of student learning.

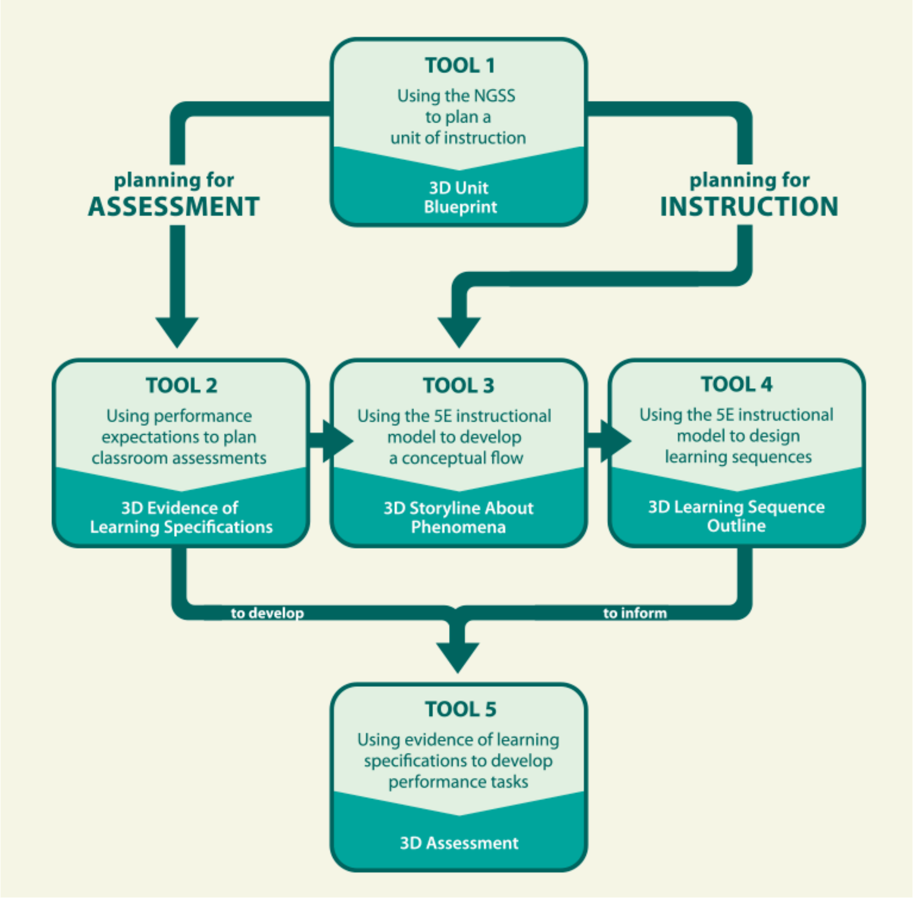

TOOL 1
Using the NGSS to plan a unit of instruction
3D Unit Blueprint
planning for ASSESSMENT
planning for INSTRUCTION
TOOL 2
Using performance expectations to plan classroom assessments
3D Evidence of Learning Specifications
TOOL 3
Using the 5E instructional model to develop a conceptual flow
3D Storyline About Phenomena
TOOL 4
Using the 5E instructional model to design learning sequences
3D Learning Sequence Outline
to develop
to inform
TOOL 5
Using evidence of learning specifications to develop performance tasks
3D Assessment

Figure 1. The Five Tools and Processes for NGSS

The Five Tools and Processes (see Figure 1) provide a concrete process for following the key recommendations within Understanding by Design (Wiggins & McTigue, 2005); mainly, that effective curriculum is planned backward from long-term desired results. Planning backward helps avoid the common problem of activity-oriented teaching (hands-on learning) in which no clear priorities and purposes are apparent or articulated. For example, Tool 2 prompts the designer to design “evidence of learning specifications” and plan embedded classroom assessments before they use the 5E Instructional model (Bybee et al., 2006) to develop a conceptual flow (Tool 3) or learning sequences (Tool 4). In Tool 5, the designer goes through the process of using the evidence of learning specifications to develop a specific, concrete performance task to be used as summative assessment. Though it may seem like to adhere strictly to Understanding by Design, the designer should design the summative task itself first, the Tools advocate for developing highly specific evidence of learning specifications, but not necessarily designing the task itself until the instructional sequences have been designed.

For this project, the curriculum was designed using the *Five Tools and Processes*. Thus, for this curriculum a 3-dimensional "Unit Blueprint" was developed, followed by 3-dimensional "evidence of learning specifications," which were based on the middle school life and earth science performance expectations described in the NGSS. The final *Disruptions in Ecosystems unit* consists of five chapters (see Table X), and there are three distinct layers of assessment embedded in the unit. First, there are classroom formative assessments that provide evidence of learning in terms of Discplinary Core Ideas, Practices, and Crosscutting Concepts. The goal of formatie assessment is to monitor student learning to provide ongoing feedback that can be used by instructors to improve their teaching and by students to improve their learning (Black & Wiliam, 2009). The instructional sequence in each chapter follows the 5E model (Bybee et al., 2006) and thus concludes with an "evaluate" task, which is also treated as a formative assessment. Finally, each chapter concludes with a Chapter Assessment, a performance task which is treated as an embedded summative assessment. The goal of summative assessment is to evaluate student learning at the end of an instructional sequence by comparing it against some standard or benchmark. Each Chapter Assessment is formative in that teachers can use the feedback on students' competency with core ideas, practices, and crosscutting concepts to inform their teaching in a following unit. However, the performance tasks can also be seen as summative task—they have associated scoring guides and sample answers, and (in theory) teachers can track learning across the instructional unit using evidence from the Chapter Assessments.

The unit as a whole follows the principle of assessment-centered teaching (DiRanna et al., 2008), which refers to a system in which assessment is integrated with science teaching in ongoing cycles of planning, instruction, assessment, and revision of instruction. Though the *Disruptions in Ecosystems* unit is a fully developed curriculum, the idea is not that teachers can simply pull out the manual and teach it. Rather, the professional development provides teachers with the tools to enact assessment-centered teaching, and thus respond to feedback they receive from students via the embedded formative assessments, "Evaluate" tasks, and summative Chapter Assessments.

Chapter Assessment tasks

This paper will discuss findings about student learning from the Chapter Assessments only (and not formative tasks or "Evaluate" tasks), which were administered to students at the conclusion of each instructional sequence. The *Disruptions in Ecosystems* unit addressed disciplinary core ideas about life and earth science and included targeted instruction in the scientific practices or modeling, constructing explanations and designing solutions, and arguing from evidence. The crosscutting concepts of patterns, energy and matter, cause and effect and stability and change were woven throughout the unit (see Table 1).

Instructional Sequences	***1. Wolves in Yellowstone***	***2. Ecosystem Models***	***3. Interactions between Populations & Resources***	***4. Zebra Mussels***	***5. Designing Solutions***
Chapter Summary	Students investigate the issue of the reintroduction of wolves to the Greater Yellowstone Ecosystem.	Students explore the effects of natural disasters on ecosystems.	Students analyze the impact of humans on commercial fisheries.	Students analyze short and long-term data on the effect of zebra mussels on the Hudson River and Great Lake Ecosystems.	Students evaluate and design solutions for environmental challenges in a variety of ecosystems.
DCIs	LS2.A Interdependent relationships in ecosystems ESS3.C	LS2.B Cycles of matter and energy transfer in ecosystems	LS2.A Interdependent relationships in ecosystems ESS3.C Human	LS2.A Interdependent relationships in ecosystems LS2.C	LS2.C Ecosystem dynamics, functioning, and resilience LS4.D

	Human impacts on Earth systems	ESS2.A Earth's materials and systems PS1.B Chemical reactions	impacts on Earth systems	Ecosystem dynamics, functioning, and resilience LS4.D Biodiversity and Resources	Biodiversity and Resources ESS3.C Human impacts on Earth systems ETS1.B Developing possible solutions
Main Science and Engineering Practices	Constructing explanations and designing solutions Engaging in argumentation from evidence	Developing and using models	Analyzing and interpreting data Constructing explanations and designing solutions Engaging in argumentation from evidence	Asking Questions Analyzing and interpreting data Constructing explanations and designing solutions Engaging in argumentation from evidence	Constructing explanations and designing solutions Engaging in argumentation from evidence
Main Crosscutting Concepts	Patterns Cause and Effect	Energy and Matter Stability and Change	Cause and Effect	Stability and Change Cause and Effect Patterns	Stability and Change Cause and Effect
Primary Performance Expectations	MS-LS2-2 Construct an explanation that predicts patterns of interactions among organisms across multiple ecosystems	MS-LS2-3 Develop a model to describe the cycling of matter and flow of energy among living and non-living parts of an ecosystem	MS-LS2-1 Analyze and interpret data to provide evidence for the effects of resource availability on organisms and populations of organisms in an ecosystem	MS-LS2-4 Construct an argument supported by empirical evidence that changes to physical or biological components of an ecosystem affect populations	MS-LS2-5 Evaluate competing design solutions for maintaining biodiversity and ecosystem services

Table 1. Summary of the Disciplinary Core Ideas (DCIs), science and engineering practices, and crosscutting concepts that are addressed in Chapters 1-5 of the *Disruptions in Ecosystems* unit.

The formative assessment tasks generally related closely to the contexts and examples used in the chapters. The end-of-chapter assessments were developed in more general contexts with the intent that they could be used with other NGSS-aligned curricula. The scenario-based assessment tasks were based directly on the performance expectations targeted in each chapter. They were drafted initially by the Lawrence Hall of Science and were accompanied by "Sample Answers" written by LHS educators. These tasks were administered to the teachers who participated in the Year 2 field test (this was the first cohort of teachers) as part of the professional development sessions. These 25 teachers completed the tasks and offered written and oral feedback. Feedback addressed (1) strength of the scenarios (e.g. discussions about whether an opposing argument in an argumentation task had enough merit to make the scenario worthy of argument) (2) the appropriateness of the length and complexity of scenarios for middle school students (3) the appropriateness of the language/wording of a scenario and (4) the clarity and helpfulness of illustrations. Teachers also offered feedback on the scoring guides (discussed below). The assessment tasks were edited based on the teachers' feedback during Summer 2015 and the edited versions were administered to students during the Fall 2015 Field Test. The full set of five chapter assessments is included in Appendix A (online).

Scoring guides

Scoring guides were developed by educators and researchers on the project and edited based on feedback from the Year 2 field test teachers. The design of scoring guides was shaped by the progressions of disciplinary core ideas, practices and crosscutting concepts described in the NGSS (Achieve, 2013) and ongoing work on the development of learning progressions for the three scientific practices addressed in the unit: modeling (Rivet & Kastens, 2012), scientific explanation (Gotwals & Songer, 2013) and argument from evidence (Osborne et al., 2016). Scoring Guides for the five chapter assessments are included in Appendix B (online).

Sample

The analysis presented in this paper is based on middle school students' written responses to Chapter Assessments 1-5. The assessments were administered during regular class time to all students of teachers in the Year 2 field test; however, only a subset of the total field test teachers participated in the research component of the project and thus submitted their student assessments to be analyzed. There were 25 middle school science teachers in the Year 2 field test; 23 completed the entire field test, which concluded in January 2016. All teachers taught grades 6-8 in New York City public middle schools. All participants were also part of "Urban Advantage," a large teacher professional development program in NYC and they applied to be part of the field test.

Five teachers volunteered to participate in the research component of the project, which consisted of classroom observations, a pre- and post-interview, submitting all their chapter assessments and collaborating with the researchers on scoring the assessments and providing feedback on the assessments and scoring guides. Two teachers opted not to continue participating in the research after realizing that their schedule did not permit the required hours of participation. A sixth teacher elected to join the research team partway through the field test; thus, the analysis of the assessments includes work from students of three teachers: Mariana (grade 8, general education class only), Kate (grade 6, general education, dual language, and integrated co-teaching classes), and Laura (grade 6, general education class only). The characteristics of the students in the classes analyzed are summarized in Table 2.

Teacher	**Grade**	**#**	**Male**	**Female**	**Asian**	**Black**	**Hispanic**	**White**	**Other**	**IEP**	**ELL**	**Free Lunch**
Mariana	8	35	23	12	2	10	22	0	0	0	0	35
Kate	6	18	8	10	0	0	18	0	0	10	2	18
	6	19	7	12	0	0	19	0	0	2	2	19
	6	22	9	13	0	0	22	0	0	0	14	22
Laura	6	21	21	0	0	13	7	0	1	0	0	13
Total		115	68	47	2	23	88	0	1	12	18	107

Table 2. A summary of the demographic information for the classes that participated in the research component of the project. # refers to the total number of students in the class. "IEP" refers to the number of students in each class with an individualized education plan. "ELL" refers to the number of students in each class who have been officially designated as English Language Learners.

Laura and Mariana completed all five chapters of the Ecological Disruptions unit; Kate completed chapters 1-4. The findings reported below are based on work submitted by teachers who were able to teach all or the majority of the unit, did not become overwhelmed by the demands of the field test, and felt comfortable submitting their students' work. Kate's scores were compared with the researchers' scores to the calculated inter-rater agreement.

Findings

We plan to conduct large-scale assessment data collection (N>200 students), scoring, and psychometric analysis in Year 3 of the field test, which will provide information about validity, reliability and fairness. In Year 2 of the field test, our goal was to collect enough data to conduct a smaller-scale analysis that would allow us to "moderate" the scoring guides (Chiu, Morell & Wilson, 2013) and revise the assessments and scoring guides to improve clarity and reduce construct-irrelevant variance. Furthermore, we were hoping to be able to analyze trends in responses across Chapter 1-5 to make some initial conclusions about the instructional sensitivity of the assessments, and the learning gains students made during the unit. Finally, we wanted to gather enough sample responses to be able to provide empirical examples within the curriculum materials such that teachers would be able to see what types of answers actual students were able to generate in response to the performance tasks. We are planning to use empirical student responses to guide activities in the professional development during Years 3 and 4.

Time and resources limited our analysis to the *explanation and argumentation tasks in Chapters 1, 3, and 4 only.* This section will summarize findings (mean scores and estimates of inter-rater agreement for items) in Chapter 1, 3, and 4 as well as present examples of student work. Table 3 summarizes each instructional sequence, over-arching performance task, the specific explanation and argument prompts, and the number of students responses that were scored and then summarized in the results below.

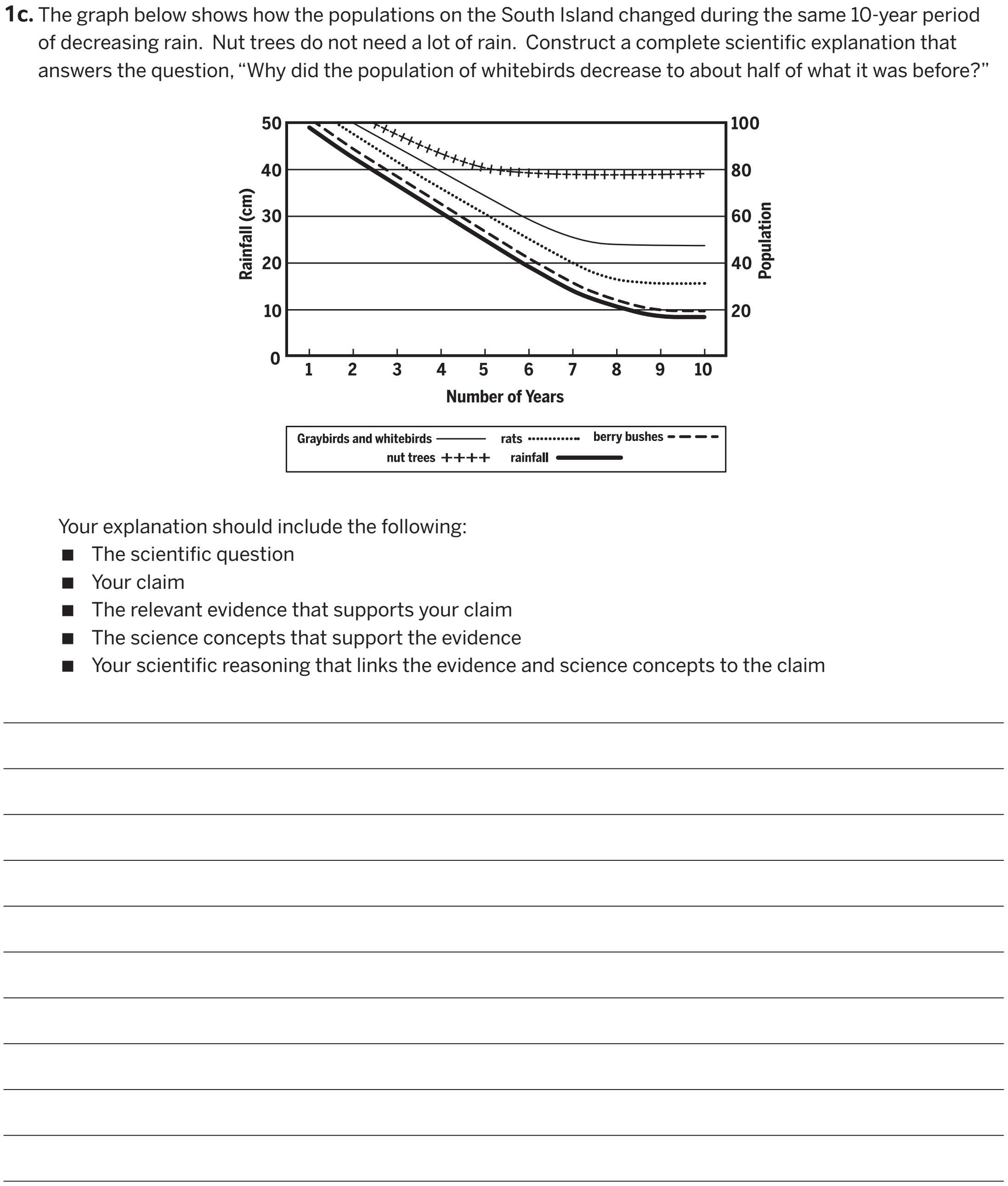
Instructional Sequence	Performance Task	Explanation and Argument Prompts	# scored responses
1. Wolves in Yellowstone	Students develop an explanation about the patterns of interactions in different ecosystems that explain the relationships among and between organisms and abiotic components of ecosystems. Interactions include competitive, predatory, and mutually beneficial relationships.	Explanation (Item 1c): Why did the population of whitebirds decrease to about half of what it was before? Argument (Items 2, 2a): Do increases in human population cause negative impacts on earth?	86
2. Ecosystem models	Students develop a model based on the conservation of matter and energy flow into and out of various ecosystems and describe interactions among biotic and abiotic factors using a food web.	NA	NA
3. Interactions between Populations & Resources	Students develop an explanation [using evidence] about the cause and effect relationships of resource availability on individuals and populations of organisms in an ecosystem.	Explanation (Item 2): Why does converting the grassland to farmland affect the deer population? Argument (Items 3, 4): Should the farmland be converted back to grassland?	87
4. Zebra Mussels	Students develop an argument using empirical evidence from their explanation showing that ecosystems are dynamic and that small changes in one part might cause big changes in another.	Argument (Items 2, 3): Can small changes to an abiotic factor in an ecosystem lead to large changes in populations?	87
5. Designing Solutions	Students develop a method for monitoring and minimizing a human impact on the environment based on understanding the conservation of matter and flow of energy, the influence of availability of resources on individuals and populations of organisms, and the dynamic nature of ecosystems, considering criteria and constraints of a problem. Competing decisions are evaluated.	Explanation (Item 1a): How does the overfishing of lobsters and snappers affect the amount of seaweed in the ecosystem? Argument (Item 1b): Why is one proposed solution more sustainable than the other?	20

Table 3. Total pool of assessments analyzed

Example 1: Assessing 'constructing a scientific explanation about ecological relationships'

The Chapter 1 assessment first asked students to describe patterns of interactions in a particular ecosystem and explain patterns on graphs of species abundances in items 1a and 1b. Then, students were asked to construct a complete scientific explanation for why, after a dry period, the number of whitebirds in the ecosystem decreased to about half of what it was before.

The item (1c) is shown in Figure 2.



1c. The graph below shows how the populations on the South Island changed during the same 10-year period of decreasing rain. Nut trees do not need a lot of rain. Construct a complete scientific explanation that answers the question, "Why did the population of whitebirds decrease to about half of what it was before?"

Rainfall (cm)
50
40
30
20
10
0
Population
100
80
60
40
20
1
2
3
4
5
6
7
8
9
10
Number of Years
Graybirds and whitebirds
rats
berry bushes
nut trees
rainfall

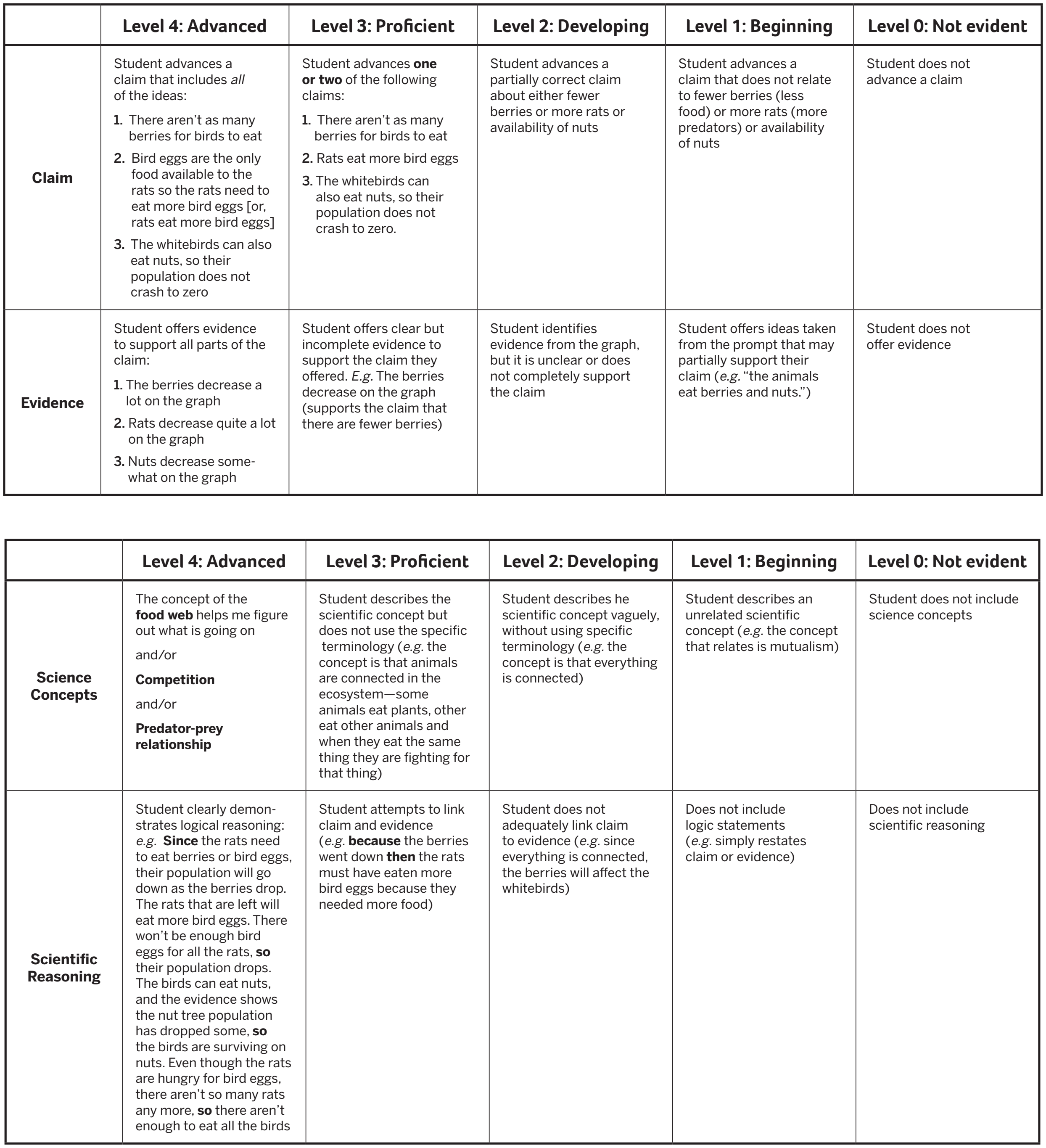
Your explanation should include the following:

- The scientific question
- Your claim
- The relevant evidence that supports your claim
- The science concepts that support the evidence
- Your scientific reasoning that links the evidence and science concepts to the claim

Figure 2. Chapter 1 performance task assessing students' ability to explain why whitebirds decrease by approximately half

A discussion of how the unit treated scientific concepts of food webs and inter-dependent relationships in ecosystems (e.g. competition and predator-prey), the practice of constructing a scientific explanation, and the crosscutting concept "patterns" is discussed elsewhere (Nagle et al., 2016). However, relevant to the findings presented here, students were introduced to the "Scientific Explanation Tool" in Chapter 1, which provided scaffolding for the processes of generating a claim, identifying data, sorting data into relevant and non-relevant evidence buckets, identifying scientific concepts that help link the claim and the evidence, and finally connecting the claim, evidence and concepts with scientific reasoning. The bulleted list below the graph uses the same language as the Explanation Tool and was added based on feedback from the Year 2 field test teachers, many of whom thought that students would need a stronger cue that they were expected to have a claim, evidence, concepts and reasoning, rather than simply providing a single sentence explanation (i.e. a claim only).

The rubric used to score the performance task above is shown below in Figure 3. Briefly, middle school students were expected to make a claim about why the birds decreased to half (mainly, that the birds has less food to eat), support the claim with evidence from the graph that was provided, and then use the concept of inter-relatedness of species, food webs, or competition (other scientific concepts were accepted) to explain that since the berries decreased, the birds also decreased, though did not crash to zero since there was still an existing food source.

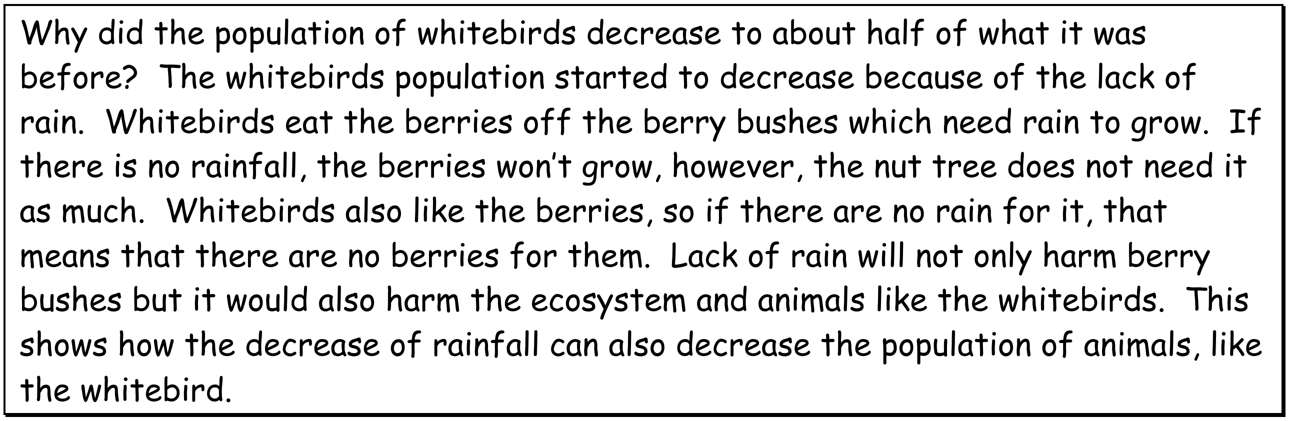


	Level 4: Advanced	Level 3: Proficient	Level 2: Developing	Level 1: Beginning	Level 0: Not evident
Claim	Student advances a claim that includes *all* of the ideas: 1. There aren't as many berries for birds to eat 2. Bird eggs are the only food available to the rats so the rats need to eat more bird eggs [or, rats eat more bird eggs] 3. The whitebirds can also eat nuts, so their population does not crash to zero	Student advances **one or two** of the following claims: 1. There aren't as many berries for birds to eat 2. Rats eat more bird eggs 3. The whitebirds can also eat nuts, so their population does not crash to zero.	Student advances a partially correct claim about either fewer berries or more rats or availability of nuts	Student advances a claim that does not relate to fewer berries (less food) or more rats (more predators) or availability of nuts	Student does not advance a claim
Evidence	Student offers evidence to support all parts of the claim: 1. The berries decrease a lot on the graph 2. Rats decrease quite a lot on the graph 3. Nuts decrease somewhat on the graph	Student offers clear but incomplete evidence to support the claim they offered. *E.g.* The berries decrease on the graph (supports the claim that there are fewer berries)	Student identifies evidence from the graph, but it is unclear or does not completely support the claim	Student offers ideas taken from the prompt that may partially support their claim (*e.g.* "the animals eat berries and nuts.")	Student does not offer evidence

	Level 4: Advanced	Level 3: Proficient	Level 2: Developing	Level 1: Beginning	Level 0: Not evident
Science Concepts	The concept of the **food web** helps me figure out what is going on and/or **Competition** and/or **Predator-prey relationship**	Student describes the scientific concept but does not use the specific terminology (*e.g.* the concept is that animals are connected in the ecosystem—some animals eat plants, other eat other animals and when they eat the same thing they are fighting for that thing)	Student describes he scientific concept vaguely, without using specific terminology (*e.g.* the concept is that everything is connected)	Student describes an unrelated scientific concept (*e.g.* the concept that relates is mutualism)	Student does not include science concepts
Scientific Reasoning	Student clearly demonstrates logical reasoning: *e.g.* **Since** the rats need to eat berries or bird eggs, their population will go down as the berries drop. The rats that are left will eat more bird eggs. There won't be enough bird eggs for all the rats, **so** their population drops. The birds can eat nuts, and the evidence shows the nut tree population has dropped some, **so** the birds are surviving on nuts. Even though the rats are hungry for bird eggs, there aren't so many rats any more, **so** there aren't enough to eat all the birds	Student attempts to link claim and evidence (*e.g.* **because** the berries went down **then** the rats must have eaten more bird eggs because they needed more food)	Student does not adequately link claim to evidence (*e.g.* since everything is connected, the berries will affect the whitebirds)	Does not include logic statements (*e.g.* simply restates claim or evidence)	Does not include scientific reasoning

Figure 3. Rubric for scoring the scientific explanation performance task 1c on the Chapter 1 assessment.

Did students in grades 6, 7 and 8 understand what they were being asked to explain? And did they produce explanations that met the expectations described in the scoring guide? Figure 4 shows an explanation constructed by Amanda, a sixth-grader in a general education classroom.

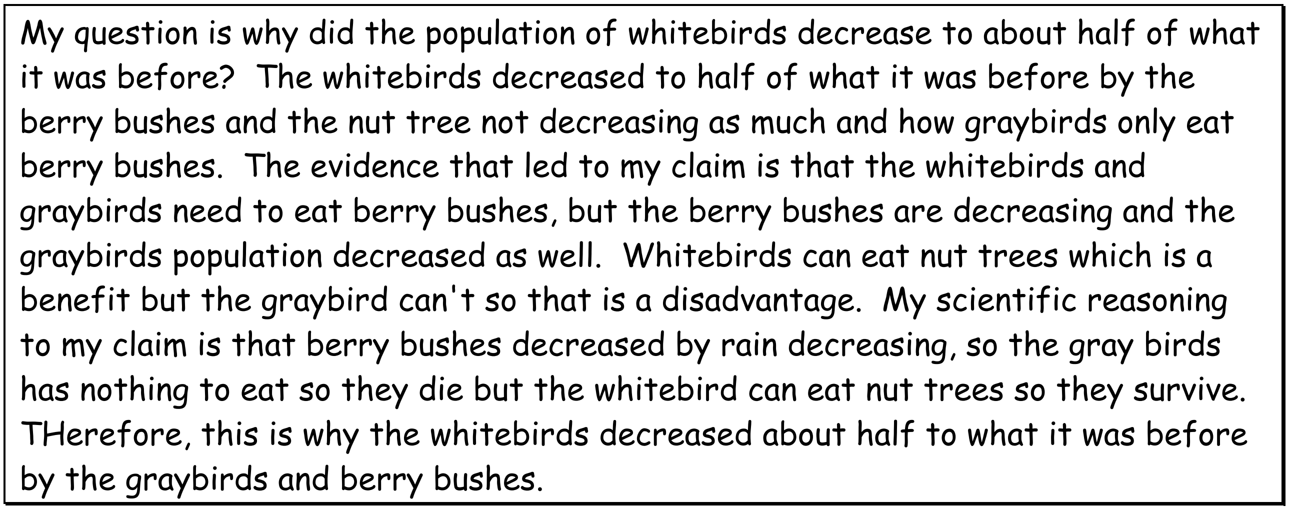


Why did the population of whitebirds decrease to about half of what it was before? The whitebirds population started to decrease because of the lack of rain. Whitebirds eat the berries off the berry bushes which need rain to grow. If there is no rainfall, the berries won't grow, however, the nut tree does not need it as much. Whitebirds also like the berries, so if there are no rain for it, that means that there are no berries for them. Lack of rain will not only harm berry bushes but it would also harm the ecosystem and animals like the whitebirds. This shows how the decrease of rainfall can also decrease the population of animals, like the whitebird.

Figure 4. Sixth-grader Amanda's response to Item 1c on the Chapter 1 assessment: "Why did the population of whitebirds decrease to about half of what it was before?" Rater 1: Claim (3), Evidence (2), Concepts (3), Reasoning (2); Rater 2: Claim (3), Evidence (2), Concepts (3), Reasoning, (2).

Amanda's explanation received 10/15 (75%)—she received full credit for her claim that "whitebirds population started to decrease because of the lack of rain. Whitebirds eat the berries off the berry bushes which need rain to grow." However, the evidence she cites is not specific enough to earn her a "Proficient" rating for evidence; mainly, she does not say, "on the graph, the berries are decreasing." In fact, her teacher provided written feedback on her assessment that she should "pull from the graph" to support her claim. She does describe the patterns of interactions among species in this ecosystem, yet does not use any specific scientific terminology to refer to the interactions, which earned a "proficient" for science concepts. Given that the response does not contain logical language that connects the claim, evidence and concepts, Amanda's reasoning was rated "developing."

Eric, also a sixth-grader, offered a response that was rated as "proficient" for claim, evidence, concepts and reasoning (Figure 5).



My question is why did the population of whitebirds decrease to about half of what it was before? The whitebirds decreased to half of what it was before by the berry bushes and the nut tree not decreasing as much and how graybirds only eat berry bushes. The evidence that led to my claim is that the whitebirds and graybirds need to eat berry bushes, but the berry bushes are decreasing and the graybirds population decreased as well. Whitebirds can eat nut trees which is a benefit but the graybird can't so that is a disadvantage. My scientific reasoning to my claim is that berry bushes decreased by rain decreasing, so the gray birds has nothing to eat so they die but the whitebird can eat nut trees so they survive. THerefore, this is why the whitebirds decreased about half to what it was before by the graybirds and berry bushes.

Figure 5: Sixth-grader Eric's response to Item 1c on the Chapter 1 assessment: "Why did the population of whitebirds decrease to about half of what it was before?" Rater 1: Claim (3), Evidence (3), Concepts (3), Reasoning (3); Rater 2: Claim (3), Evidence (3), Concepts (3), Reasoning, (3).

The main difference between Eric's response and Amanda's is that Eric specifically references the decreasing berries in the graph. However, even Eric does not cite numbers from the graph, which begs the question why did students avoid citing highly specific evidence to support their claims, despite having this aspect of explanation included on the "Explanation Tool," and having several opportunities to practice this aspect of constructing explanations in Chapter 1?

Amanda and Eric constructed explanations that, for the most part, met the expectations the teachers and researchers had for middle school students in Chapter 1. There were examples of explanations that exceeded the expectations by including a more complex claim (e.g. that the whitebirds decreased to half because they had less to eat and rats were more likely to eat their eggs, but did not disappear completely because they still had nuts as a food source) as well as highly specific vocabulary. In addition, there were many students who struggled to meet the expectations. For example, Alfredo's explanation (shown in Figure 6) provides evidence that he was not thinking about the patterns of interactions between species in this ecosystem. When asked why the whitebirds decreased during this dry period, he said, because "rain is life" and without rain “all populations will decreased [sic].” This explanation shows that Alfredo understood he was being asked (mainly, why were there fewer birds?), but did not engage in explanation in the way in which students rehearsed this practice with the Explanation Tool in Chapter 1 and did not have a strong enough grasp on the idea that species interact in an ecosystem, and that the lack of rain directly affects the berries and thus indirectly affects the population of whitebirds.

> The reason why the population decreased a lot is because not many of the other birds had enough rain/water. Rain is life because rain contains things such as life and in support life and other populations in the ecosystem. Therefore without rain all populations will decreased. In the graph it shows a lot of decretion towards other things in the community.

Figure 6. Sixth-grader Alfredo’s response to the Chapter 1 explanation task

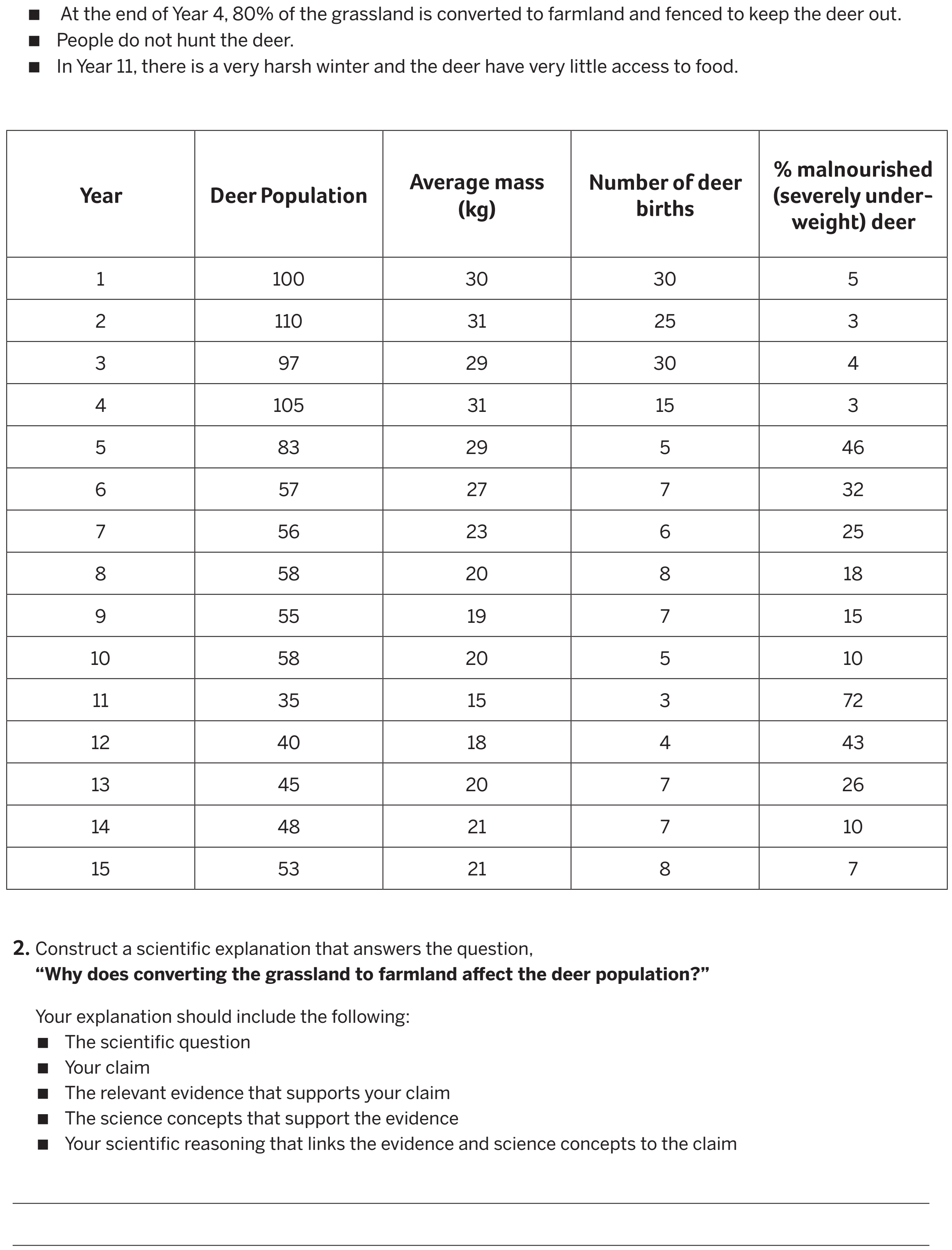
Jairo’s explanation (Figure 7) also provides evidence that some students struggled to construct an explanation that was consistent with the data provided.

> I think the population of white birds decrease to about half of what it was before because since it rained a lot many of them moved to a different place. I claimed this has happen because of the amount of rainfall. The evidence I have is that since in the years it rained the birds left so if it rains the birds can't fly so they need to be dry in order to fly.

Figure 7. Sixth-grader Jairo’s response response to the Chapter 1 explanation task

The responses above provide evidence that students understood the prompt, and understood what they were being asked to do. In general, students tended not to cite specific evidence from the graph, perhaps because they didn’t see it as necessary to make their point? Kate provided written feedback to many students on the Chapter 1 assessments to pull information from the graph. The mean score on evidence was 1.61, which is below developing, The unexpectedly low scores on evidene beg the question – why did students rely on generalities and not cite specific evidence from the graph? Why did they not see this as necessary? And did they get any better at this as the unit progressed?

Students' ability to construct an explanation was next formally assessed on Item 2 in the Chapter 3 assessment. In some ways, this explanation was structurally and conceptually similar to the explanation students were asked to construct in Chapter 1. In Item 2 on the Chapter 3 assessment, students were asked why a change to the environment resulted in a change to a population. However, whereas the disciplinary core idea students were expected to develop in their answer was inter-dependent relationships in Chapter 1, in Chapter 3 students were expected to discuss "resource availability and effects on individuals and populations of organisms in an ecosystem." In terms of crosscutting concepts, the Chapter 1 explanation task expected students to draw on their understanding of patterns of interactions; the Chapter 3 explanation task expected students to rely more strongly on the concept of cause and effect. Fitting the data to variants of the basic Rasch model, planned for Year 3, may be able to shed light on the effect of the *context* of the task on students' ability to argue. In this analysis, we simply ask the question: Did students' ability to construct an explanation seem to improve (keeping in mind that the concepts were different, and thus item difficulty could have affected out ability to detect growth). Furthermore, what were the qualitative differences between the explanations offered in Chapter 1 versus Chapter 3? Item 2 from the Chapter 3 assessment is shown in Figure 8.



- At the end of Year 4, 80% of the grassland is converted to farmland and fenced to keep the deer out.
- People do not hunt the deer.
- In Year 11, there is a very harsh winter and the deer have very little access to food.

Year	Deer Population	Average mass (kg)	Number of deer births	% malnourished (severely underweight) deer
1	100	30	30	5
2	110	31	25	3
3	97	29	30	4
4	105	31	15	3
5	83	29	5	46
6	57	27	7	32
7	56	23	6	25
8	58	20	8	18
9	55	19	7	15
10	58	20	5	10
11	35	15	3	72
12	40	18	4	43
13	45	20	7	26
14	48	21	7	10
15	53	21	8	7

2. Construct a scientific explanation that answers the question,
"Why does converting the grassland to farmland affect the deer population?"

Your explanation should include the following:

- The scientific question
- Your claim
- The relevant evidence that supports your claim
- The science concepts that support the evidence
- Your scientific reasoning that links the evidence and science concepts to the claim

Figure 8. Item 2 on Chapter 3 assessment: Construct an explanation that answers, "Why does converting the grassland to farmland affect the deer population?"

> Why does converting the grassland to farmland affect the deer population?
> Converting the grassland to farmland affects the deer population because then they would have less food and it would be harder to get food. The chart shows that year 11 was the deer most decreased population. The chart also shows that there was a decrease of the amount of deer births. The highest amount of deer births was in year 1 and 3 and the amount was 30. The lowest amount was in year 11, the amount of deer births was 3. In addition, the weight of the deer's were really severe. The deers were starving. Science concepts to support my evidence and claim is that the deer population started to decrease once the grassland was being converted to farmland.

Figure 9. Sixth-grader Amanda's explanation for, "Why does converting the grassland to farmland affect the deer population? Rater 1 scores: Claim (3), Evidence (3), Concepts (3), Reasoning (2); Rater 2 scores: Claim (3), Evidence (3), Concepts (1), Reasoning (2)

Compared to her Chapter 1 explanation, sixth-grader Amanda seems to have taken her teacher's feedback and offers more specific evidence to support her claim that converting the grassland meant less food for the deer. Furthermore, the logic is partially there—e.g. the deer started decreasing after the conversion because they were starving. However, the scientific concept of resource availability is not connected to the claim or evidence

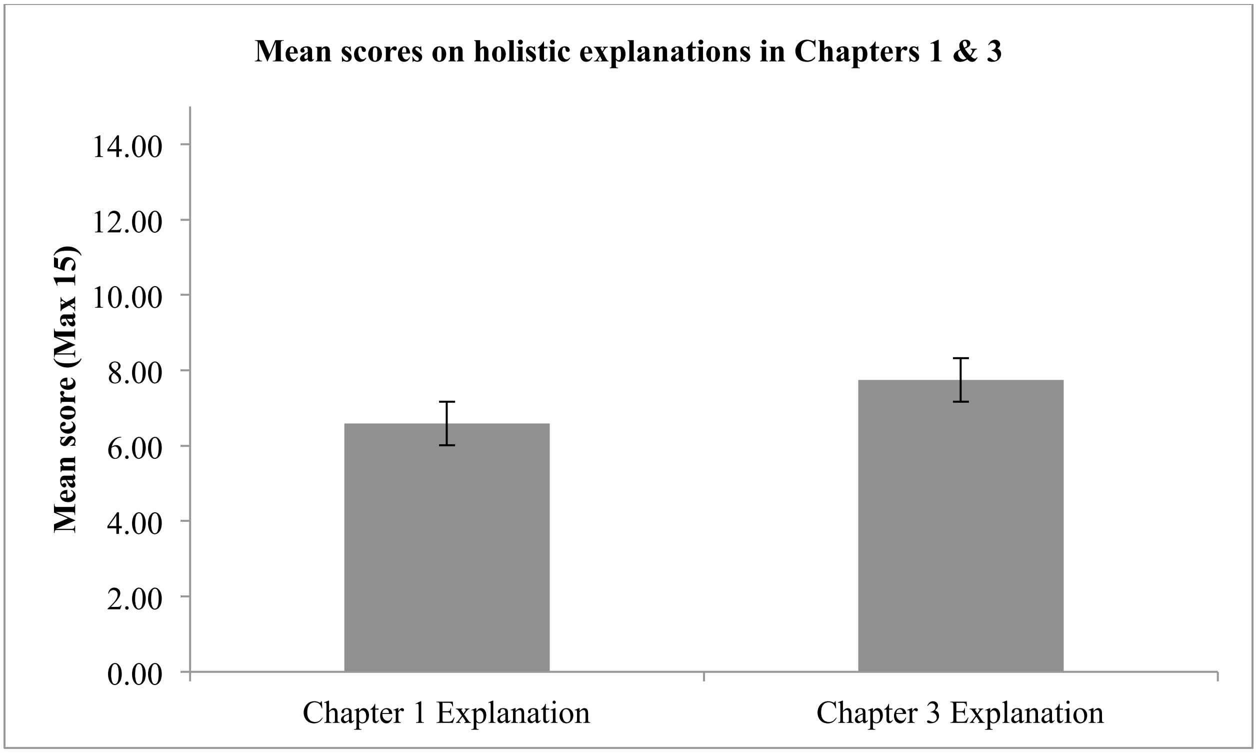


Figure 10. Mean scores on holistic explanations in Chapter 1 & 3

Figure 10 compares mean scores on the full explanation item between Chapter 1 and Chapter 3. As mentioned above, it may not be appropriate to compare mean scores to measure growth because features of each task may affect item difficulty. However, examination of the tasks themselves and the sample of student responses do not suggest a dramatic difference in difficulty. Rather, it appears that, in general, students improved (slightly) at constructing explanations, though not dramatically. Furthermore, mean scores on claims, evidence, concepts and reasoning were still below "proficient" in Chapter 3, meaning that these tasks do not appear to suffer from a ceiling effect.

Mean Scores (Full score)	Chapter 1	Inter-rater agreement	Chapter 3	Inter-rater agreement
Claim (3)	2.17	.90	2.40	.89
Evidence (4)	1.61	.89	1.75	.86
Scientific Concepts (4)	1.37	.85	1.75	.80

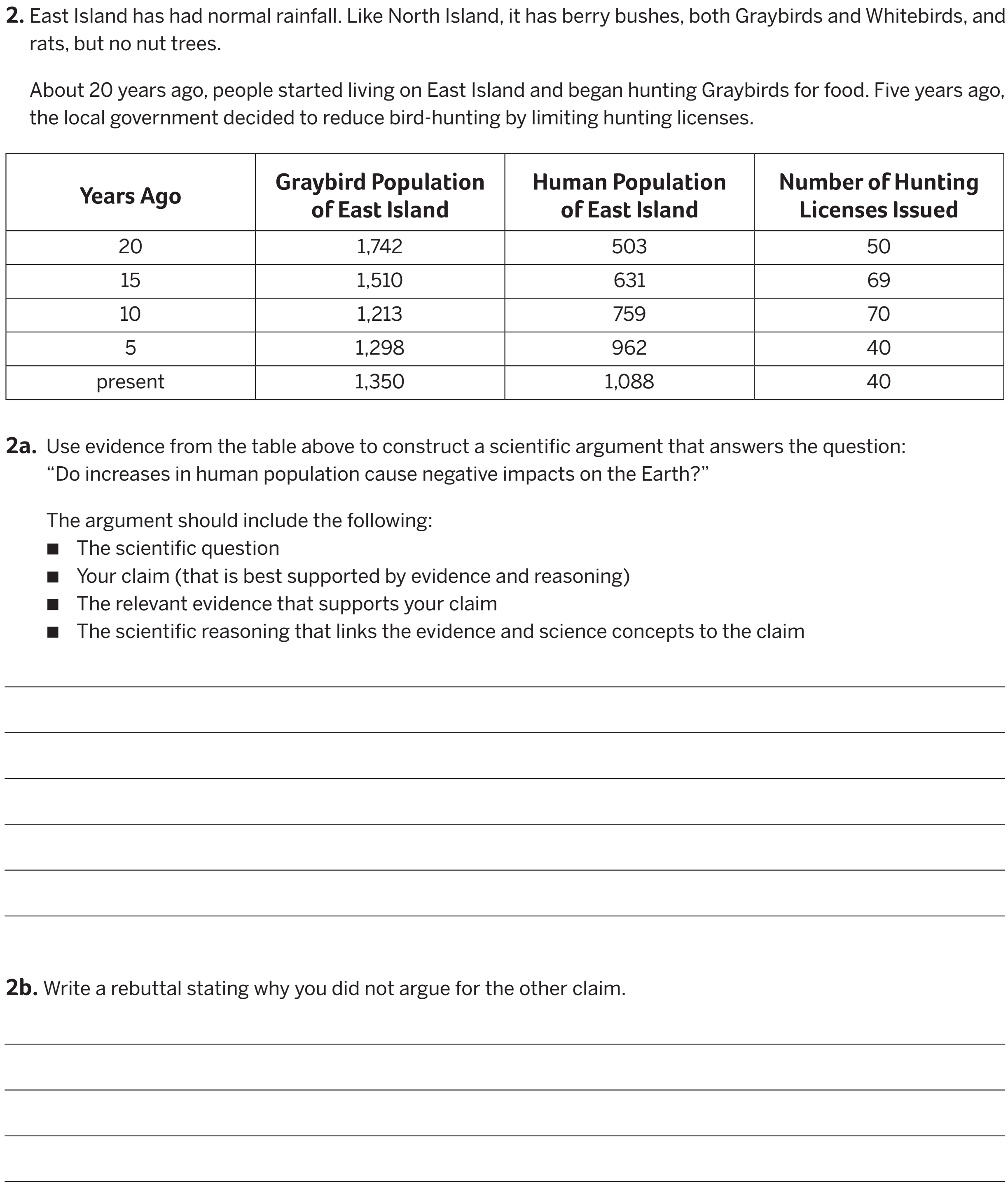
Scientific Reasoning (4)	1.42	.78	1.89	.75
Holistic (15)	6.58		7.74	

Table 4. Summary of mean scores and inter-rater agreement for explanation tasks in Chapters 1 and 3

Column 5 of Table 4 shows inter-rater agreement, which in this case is simply the percentage of times the raters landed on the same score (so, doesn't correct for chance and does not take into account the ordered categories—more sophisticated measures planned for Years 3-4). In general, raters agreed about scores for Claim and Evidence and agreed less about scores on Reasoning. This is perhaps not surprising, given that even the researchers had trouble identifying and characterizing strong reasoning; there is work to be done on the scoring guide and Explanation Tool to clarify reasoning.

Example 2: Assessing 'arguing from evidence about human impact on ecosystems'

Students were assessed on their ability to argue from evidence in Chapters 1, 3, 4 and 5; however, only Chapters 1, 3, and 4 are discussed below. In the argumentation tasks, the Claim and Evidence parts of the argument were identical to the explanations students constructed in the chapters. However, the reasoning was structured differently in the Argument Tool; whereas in an explanation, students were supported in connecting their claim and evidence with scientific concepts and logical reasoning, in an argument students were supported in reasoning about the strength of the evidence, a difference that will be discussed in the Discussion. Furthermore, for argumentation tasks, students were asked to construct a written rebuttal, or to explain why they didn't argue for the alternative claim. Item 2 from the Chapter 1 assessment is shown



2. East Island has had normal rainfall. Like North Island, it has berry bushes, both Graybirds and Whitebirds, and rats, but no nut trees.

About 20 years ago, people started living on East Island and began hunting Graybirds for food. Five years ago, the local government decided to reduce bird-hunting by limiting hunting licenses.

Years Ago	Graybird Population of East Island	Human Population of East Island	Number of Hunting Licenses Issued
20	1,742	503	50
15	1,510	631	69
10	1,213	759	70
5	1,298	962	40
present	1,350	1,088	40

2a. Use evidence from the table above to construct a scientific argument that answers the question: "Do increases in human population cause negative impacts on the Earth?"

The argument should include the following:

- The scientific question
- Your claim (that is best supported by evidence and reasoning)
- The relevant evidence that supports your claim
- The scientific reasoning that links the evidence and science concepts to the claim

____

____

____

____

____

____

2b. Write a rebuttal stating why you did not argue for the other claim.

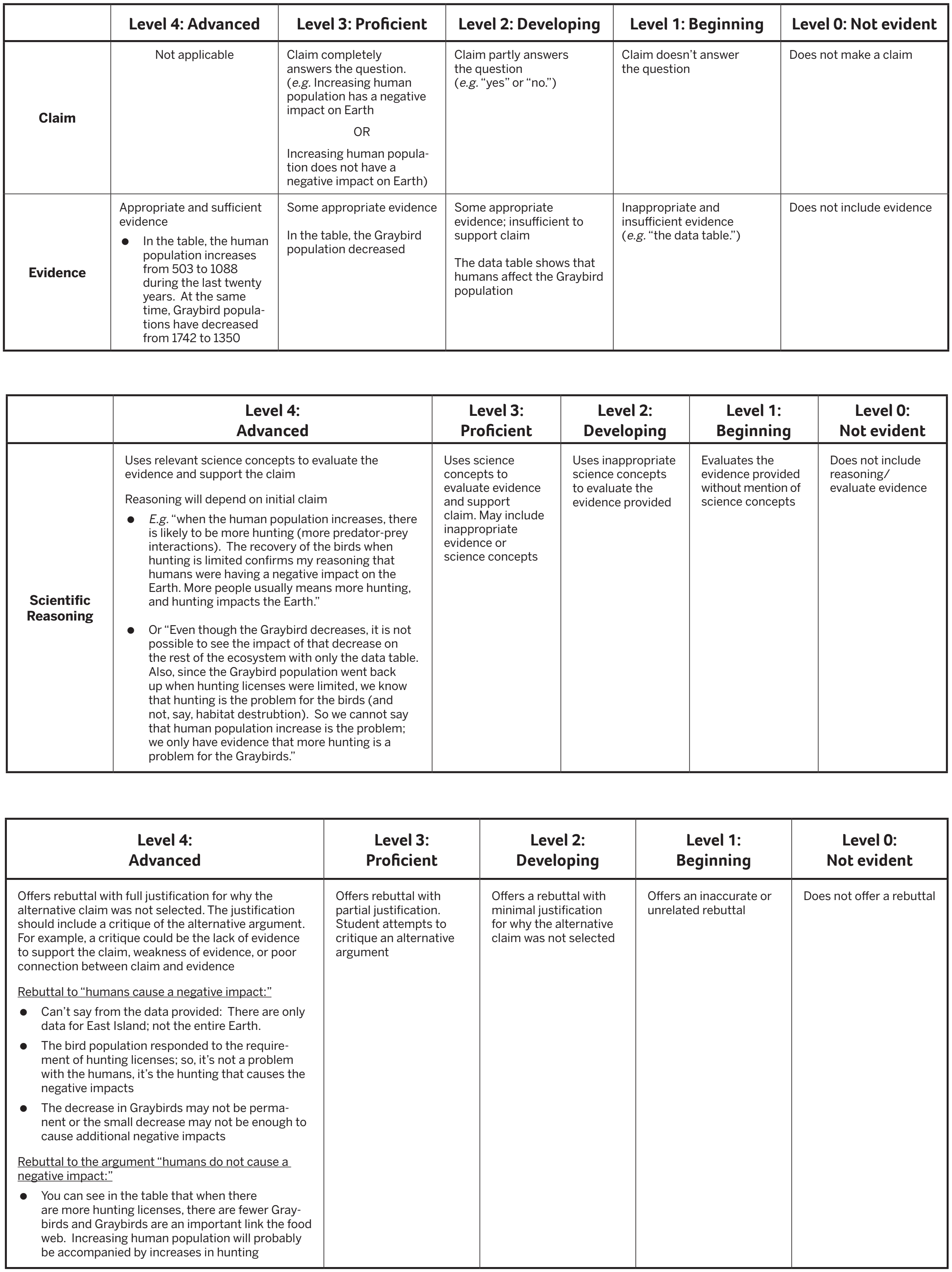
____

____

____

____

Figure 11. Item 2 from the Chapter 1 assessment



	Level 4: Advanced	Level 3: Proficient	Level 2: Developing	Level 1: Beginning	Level 0: Not evident
Claim	Not applicable	Claim completely answers the question. (*e.g.* Increasing human population has a negative impact on Earth OR Increasing human population does not have a negative impact on Earth)	Claim partly answers the question (*e.g.* "yes" or "no.")	Claim doesn't answer the question	Does not make a claim
Evidence	Appropriate and sufficient evidence ● In the table, the human population increases from 503 to 1088 during the last twenty years. At the same time, Graybird populations have decreased from 1742 to 1350	Some appropriate evidence In the table, the Graybird population decreased	Some appropriate evidence; insufficient to support claim The data table shows that humans affect the Graybird population	Inappropriate and insufficient evidence (*e.g.* "the data table.")	Does not include evidence

	Level 4: Advanced	Level 3: Proficient	Level 2: Developing	Level 1: Beginning	Level 0: Not evident
Scientific Reasoning	Uses relevant science concepts to evaluate the evidence and support the claim Reasoning will depend on initial claim ● *E.g.* "when the human population increases, there is likely to be more hunting (more predator-prey interactions). The recovery of the birds when hunting is limited confirms my reasoning that humans were having a negative impact on the Earth. More people usually means more hunting, and hunting impacts the Earth." ● Or "Even though the Graybird decreases, it is not possible to see the impact of that decrease on the rest of the ecosystem with only the data table. Also, since the Graybird population went back up when hunting licenses were limited, we know that hunting is the problem for the birds (and not, say, habitat destrubtion). So we cannot say that human population increase is the problem; we only have evidence that more hunting is a problem for the Graybirds."	Uses science concepts to evaluate evidence and support claim. May include inappropriate evidence or science concepts	Uses inappropriate science concepts to evaluate the evidence provided	Evaluates the evidence provided without mention of science concepts	Does not include reasoning/ evaluate evidence

Level 4: Advanced	Level 3: Proficient	Level 2: Developing	Level 1: Beginning	Level 0: Not evident
Offers rebuttal with full justification for why the alternative claim was not selected. The justification should include a critique of the alternative argument. For example, a critique could be the lack of evidence to support the claim, weakness of evidence, or poor connection between claim and evidence Rebuttal to "humans cause a negative impact:" ● Can't say from the data provided: There are only data for East Island; not the entire Earth. ● The bird population responded to the requirement of hunting licenses; so, it's not a problem with the humans, it's the hunting that causes the negative impacts ● The decrease in Graybirds may not be permanent or the small decrease may not be enough to cause additional negative impacts Rebuttal to the argument "humans do not cause a negative impact:" ● You can see in the table that when there are more hunting licenses, there are fewer Graybirds and Graybirds are an important link the food web. Increasing human population will probably be accompanied by increases in hunting	Offers rebuttal with partial justification. Student attempts to critique an alternative argument	Offers a rebuttal with minimal justification for why the alternative claim was not selected	Offers an inaccurate or unrelated rebuttal	Does not offer a rebuttal

Figure 12. Scoring guide for the argument students construct on the Chapter 1 assessment.

Figure 13 shows sixth-grader Eric’s argument. Eric’s argument was chosen as an example because it has features that were common among many student responses. First, he states a claim that answers the question and cites evidence from the table. On average, students tended to cite relevant evidence more frequently on the argument items than on the explanation items (e.g. in Chapter 1, the mean explanation-evidence score was 1.61 and the mean argument-evidence score was 2.19).

> Do increases in human population cause negative impacts on the Earth. Human population do have negative impacts on the Earth, when they are increasing. The evidence that supports my claim is that over the last 20 years 269 graybirds have been hunted when human population increases and it is a big reduce on a small island. Therefore, this is why humans causes big impacts on Earth by reducing lots of population. I did not argue for the other claim which is humans not having negative impacts because it doesn't have enough evidence to prove it and the claim is not as strong as humans harming Earth.

Figure 13. Sixth-grader Eric’s argument answering the question, “Do increases in human population cause negative impacts on the Earth?”

Despite students being more likely to cite relevant evidence in their arguments, there were still quite a few students who offered an opinion without evidence or reasoning, such as eighth-grader Alice (Figure 14).

> Sometimes it really can. The reason why I say that is because a lot of people out there don't recycle and then it makes our earth a dirty place. The reason I didn't argue for this claim is because it's true that most humans don’t take care of our community

Figure 14. Eighth-grader Alice’s argument answering the question, “Do increases in human population cause negative impacts on the Earth?”

Despite students, in general, performing closer to ‘proficient’ on the Chapter 1 argument than on the explanation, it is hard to shake the feeling that this item was simply a “gimme.” Even by the time they have reached the end of Chapter 1, students have already heard the idea that humans have an impact on Earth; indeed, they have likely been hearing this message for

some time by the time they are in middle school. So, it seems like there may not be very much cognitive weighing of alternatives; rather, students simply need to find the place in the data table from which to pull the numbers. Furthermore, the quality of the written rebuttals was low, though teachers rated the rebuttals highly (the inter-rater agreement for rebuttals is quite low, see Table 5). The teacher tended to award ‘proficient’ when students simply said, generally, “I didn’t argue for the other side because there wasn’t enough evidence.” The researchers were looking for a deeper discussion of why the evidence does not support the other argument; however, admittedly, this would be a tough rebuttal to mount with this particular task. Implications for the design of argumentation tasks will be addressed in the discussion.
Looking at arguments across the chapters one can see that students did better in Chapter 3, but worse in Chapter 4. In Chapter 4 they were given 3 different data sets, and had to synthesize a claim, so this additional complexity likely contributed to high difficulty.

Figure 15 compares students’ performance on the summative argumentation tasks in Chapters 1, 3, and 4. Whereas performance improves, on average, between Chapters 1 and 3, performance dipped in Chapter 4. The Chapter 4 task was significantly more complex, however, and students had to evaluate three separate data tables and synthesize an overall effect. This made the task more cognitively demanding, but also a more authentic argumentation scenario—there were arguments to be made for both sides. However, despite the rich, authentic context, students still tended to offer surface-level rebuttals without engaging with the scientific content (e.g. “there isn’t enough evidence for the other claim”).

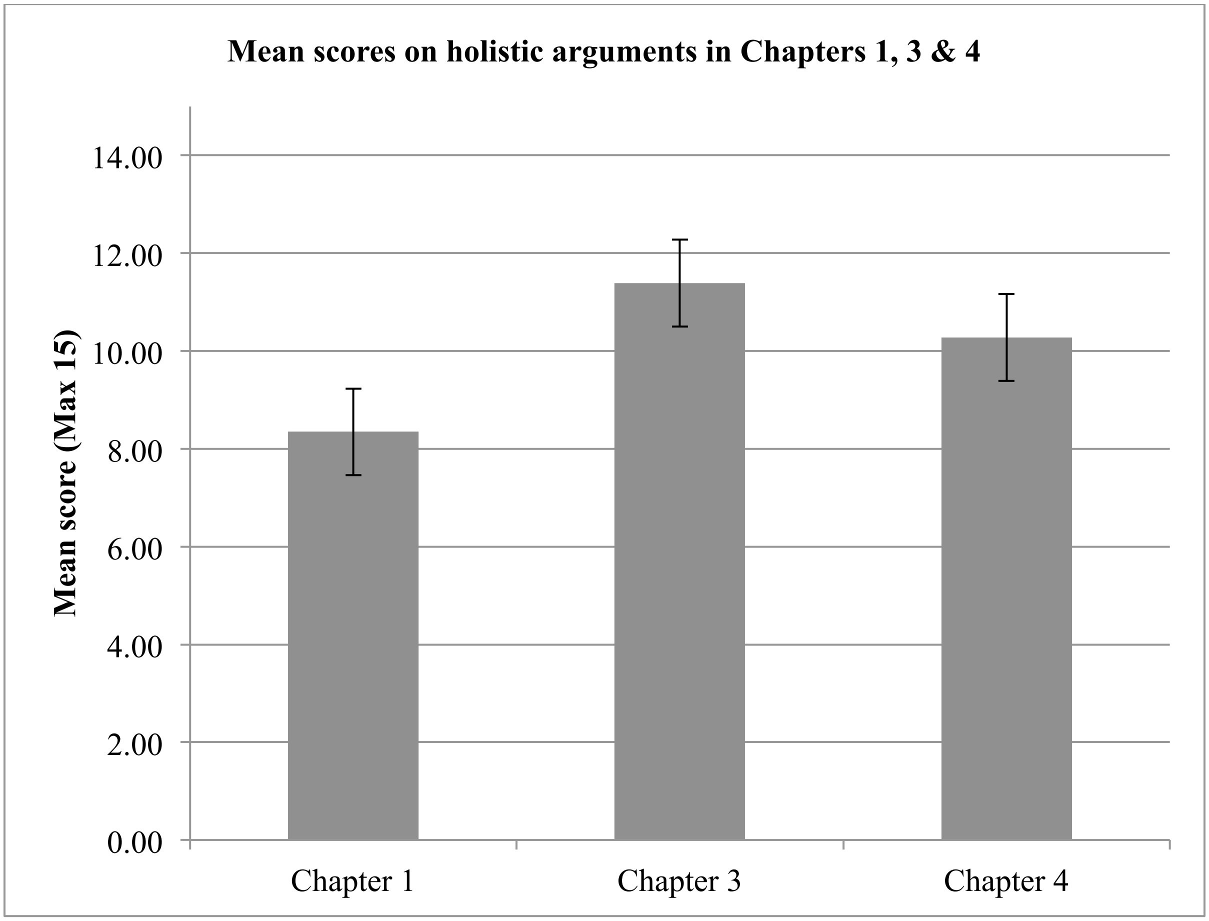


Figure 15. A comparison of mean scores on argumentation tasks across Chapters 1, 3, and 4.

Mean Scores (Full score)	Chapter 1	Inter-rater agreement	Chapter 3	Inter-rater agreement	Chapter 4	Inter-rater agreement
Claim (3)	2.43	.95	3.06	.96	2.89	.95
Evidence (4)	2.19	.85	2.50	.86	2.39	.87
Scientific Reasoning (4)	1.75	.80	2.56	.78	2.28	.77
Rebuttal (4)	2.00	.45	3.28	.40	2.72	.66
Holistic (15)	8.35		11.39		10.28	

Table 5. Summary of mean scores and inter-rater agreement for argumentation tasks in Chapters 1, 3 and 4

Discussion

The findings presented in this paper have implications for (1) the revision of the assessments and scoring guides themselves (2) design considerations for the next year of this study (3) the revisions of the curriculum and professional development and (4) the development of responsible NGSS-aligned assessment broadly.

- **Conclusion #1**: Evidence for cognitive validity

Middle school students who completed the *Disruptions in Ecosystems* curriculum unit seemed to know what we were asking them to do on the argumentation and explanation tasks. There were some students who demonstrated proficiency, a larger portion that demonstrated near proficiency, and some students that were still developing proficiency. But the fact that most students attended to the task and gave responses in line with the expected knowledge and practices offers baseline evidence of cognitive validity. Furthermore, there is some (very preliminary) evidence that students developed their ability to construct explanations and arguments across the unit. However, the validity case was stronger for some tasks and not for others. For example, the argumentation task that asked students to construct an argument and a rebuttal about whether humans cause negative effects on ecosystems should be revisited to see if there are ways to introduce *genuine difference* into the scenario that would prompt students to engage in weighing of alternatives, and not simply seek evidence to support a conclusion that already held.

- **Conclusion #2**: Students struggled to cite relevant evidence

Students scored unexpectedly low on citing relevant evidence. Whereas in previous research, citing relevant evidence was typically harder than stating a claim but easier than connecting claim and evidence with reasoning or critiquing a claim or evidence (MacPherson, 2015;

Osborne, et al., 2016), in this work, scores for evidence were often lower than for scientific concepts, scientific reasoning, or rebuttals. Examining the responses to the explanation and argumentation tasks indeed revealed that many students tended to completely omit reference to graphs, tables, or other data presented in the prompt. This finding warrants further investigation and possible changes to the PD and curriculum. In the PD in the Year 2 field test, it became clear that there was disagreement among teachers about whether textual evidence counts as evidence in an explanation or an argument. The PD emphasized that science evidence is not just something read in a text; rather, it is qualitative or quantitative data that we are using as justification for a claim. The data can be gathered by students or accessed from data sources or presented in a prompt; subsequently, the data becomes evidence when it is used in service of an explanation or argument (Hand, 2008). In other words, we thought it was clear in the curriculum, in the teacher's guide, and in the activities we did in the PD, that students should be pulling data into their explanations and arguments. However, few students cited data points or empirical observations to support their claims. Thus, we may need to examine how we are define evidence in the PD and curriculum materials and how we can support teachers in engaging students in productive talk and writing about evidence.

- **Conclusion #3**: Prompts for rebuttals generated surface-level critiques of the alternative claim. Teachers tended to score rebuttals higher than researchers.

Qualitative examination of the rebuttals students generated in their arguments (prompted by the question, "Why didn't you choose the other claim?") revealed that, for the most part, middle school students did not engage in thoughtful critique of the alternative. Rather, they cited a surface-level critique, such as, "The other claim is not supported by evidence." And, though such surface-level critique may be true, such a critique does not provide evidence of scientific mastery because the critique is too generally. What we are most interested in assessing in this prompt is whether the student can identify why the evidence is not as strong or which logical fallacy has been committed. Previous work has asked students, “What is the problem with the other argument?” with the goal of specifically prompting critique (MacPherson, 2015; Henderson et al., 2015; Osborne, et al., 2016). However maybe what we are really after is, “What would you say to the person that said [the opposite claim.] Construct a *counter-argument* that identifies the problem with the other person’s argument.” Cognitive interviews and piloting would be needed to hone in on which prompt is most successful. Furthermore, empirical examples of ‘advanced’ and ‘proficient’ rebuttals should help teachers and researchers be more aligned in their scoring of these items.

- **Conclusion #4**: Scientific reasoning needs to be operationalized more clearly in the curriculum and assessments

The reasoning that students produced on the assessments, as well as the feedback from teachers on the Argument and Explanation Tools and assessment tasks, indicates that students and teachers (and researchers?) are still pretty confused about scientific reasoning. In the current scoring guides, scientific concepts are separated from logical reasoning. This decision was made based on previous findings from teachers using the Developing Scientific Explanations Tool (DSET) in the Urban Advantage Program in NYC. In this previous work, teachers struggled with using relevant science concepts to help make sense of the and also using logical connection language (the language of reasoning) to explicitly connect evidence to claim. Thus, the two aspects of reasoning are kept separate on the DSET. However, even separating out reasoning, and asking students to be explicit about the reasoning that connects claim and evidence is problematic. Students often engage in reasoning naturally; however, when they are asked to explicitly “state their reasoning” they become confused because they feel like the reasoning is unable to be separated from the argument as a whole (Osborne, et al., 2016). How can we better support scientific reasoning in explanation and argumentation, such that does not seem like an unnatural act? Further investigation of the responses collected for this project is warranted; currently, it is not clear whether the change needs to happen at the curriculum, assessment, or professional development level. It seems possible that changing the prompt for the reasoning might move students in the right direction. We have discussed, for example, possibly changing the prompt to "Explain your evidence," in the hope that this might prompt students to more naturally connect claim with evidence. However, cognitive interviews and additional piloting would be needed to support such a change.

- **Conclusion #5**: Responsible assessment of student learning may require multiple modalities

All of the assessments for the unit required extensive writing, except for two modeling questions that, instead, asked students to indicate something on a picture. One (perhaps unintended?) consequence of this design decision, which was made mostly for logistical reasons, was that teachers and students completed most of the formative assessments using the Explanation and Argument tools, meaning that a large portion of class time was devoted to writing up the Tools and full explanations and arguments. However, these practices occur in other modalities, and so there will be changes to the curriculum, professional development and assessments to encourage more discussion. The feedback from the field test teachers was that there was too much writing in the curriculum (Nagle et al., 2016) and that students became disenchanted with explaining and arguing. We want our assessments to have strong features of measurement instruments such as reliability and fairness; but we also do not want to lose opportunities for students to demonstrate proficiency in different modalities. Additional analysis from the Dual Language and Integrated Co-Teaching classes may inform this work.

Our first draft of an assessment system had some strengths, but it also had some weaknesses. We plan to use the remaining two years of the project to revise the assessments and conduct more rigorous analysis such that we can support (or refute) these initial conclusions. As Dolores, a sixth-grader, lamented on her Chapter 1 assessment reflection: " I am not at mastery yet ☹." Neither are we Dolores! But we will use this first round of feedback to edit the assessments and refine our ideas about responsible NGSS-aligned assessment.

References

Achieve (2013). Next Generation Science Standards. http://www.nextgenerationscience.org

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

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258–267. doi:10.3102/0013189X07306523

McNeill, K. L. & Berland, L. (2015, April). *Design heuristics to enable students productive use of evidence in k-12 classrooms.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Chicago, Il.

Bybee, R. W. (1997). Achieving scientific literacy: From purposes to practices. Portsmouth, NH: Heinemann.

Bybee, R. W., Taylor, J., Gardner, A., Van Scotter, P. Carlson, J., Westbrook, A., & Landes, N. (2006). *The BSCS 5E instructional model: Origins, effectiveness, and applications.* Colorado Springs, CO: BSCS.

Black, P., & WIliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80, 139-148.

Chiu, T., Morell, L., & Wilson, M. Coding Student Responses: An iterative and construct-driven approach that uses multiple sources of data. Paper presented at the 2013 American Educational Research Association Annual Meeting, San Francisco, CA.

DiRanna, K., Osmundson, E., Topps, J., Barakos, L., Gearhart, M., Cerwin, K., Carnahan, D., & Strang, C. (2008). Assessment-Centered Teaching: A Reflective Practice. Corwin Press, Thousand Oaks, CA.

Gotwals, A. W., & Songer, N. B. (2013). Validity Evidence for Learning Progression-Based Assessment Items That Fuse Core Disciplinary Ideas and Science Practices. *Journal of Research in Science Teaching*, n/a–n/a. http://doi.org/10.1002/tea.21083

Hand, B. (2008). *Questions, Claims, and Evidence: The Important Place of Argument in Children's Science Writing*. Heinemann Educational Books.

Hannaway, J., & Hamilton, L. (2008). *Performance-based accountability policies: Implications for school and classroom practices*. The Urban Institute and RAND Corporation.

Validity. (1989). In R. Linn, *Educational Measurement* (pp. 137–149).

National Research Council. (2012). A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC.: Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education.

National Research Council, (2014). Developing assessments for the Next Generation Science standards. Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education, James W. Pellegrino, Mark, R. Wilson, Judith A. Koenig, and Alexandra S. Beatty, *Editors*. Division of Behavioral and Social Sciences and Education. Washington, DC.: National Academy Press.

Osborne, J., Henderson, J. B., MacPherson, A., Wild, A., Szu, E., & Yao, S-Y (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching.*

Rivet, A. E., & Kastens, K. A. (2012). Developing a construct-based assessment to examine students' analogical reasoning around physical models in Earth Science. *Journal of Research in Science Teaching*, *49*(6), 713–743.

Shavelson, R. J. & Baxter, G. P. (1991). Linking assessment with instruction. In Oser, F.K, Dick, A. & Patry, J. (Eds.), *Effective and Responsible Teaching* (80-90), San Francisco: Jossey Base Publishers.

Sunal, D. W., & Wright, E. (2006). *The Impact of State and National Standards on K-12 Science Teaching*. IAP.

Weiss, I. R., Pasley, J. D., Smith, P. S., Banilower, E. R., & Heck, D. J. (2003). *Looking inside the classroom: A study of K-12 mathematics and science education in the United States*. Horizon Research, Incorporated.

Wiggins, G., & McTighe, J. (2005). Understanding by design. Alexandria, VA: Association for Supervision and Curriculum Development.

Wiliam, D. (2010). What counts as evidence of educational achievement? The role of constructs in the pursuit of equity in assessment. *Review of Research in Education, 34*(1), 254–284. doi:10.3102/0091732X09351544