HIGH-ADVENTURE SCIENCE
Final Report
July 2012

PARTICIPANTS

1. What people have worked on your project?

<u>Concord Consortium Staff</u>

Amy Pallant, Daniel Damelin, Nathan Kimball, Robert Tinker, Sarah Pryputniewicz, Rachel Kay, Stephen Bannasch, Alex Bean, Scott Cytacki, Adam Knochowski, Ethan McElroy, Cynthia McIntyre, Noah Paessel

2. What other organizations have been involved as partners?

None

3. Have you had other collaborators or contacts?

Collaborators

Dr. Hee-Sun Lee	University of California, Berkeley
Andy Reichsman	Ames Hill Film and Video Productions
Dr. Mark Chandler	NASA Goddard Institute for Space Studies
Dr. Holly Michael	Professor of Hydrogeology, University of Delaware
Ted Sicker	NOVA
Dr. Daniel Schrag	Harvard University
Dr. Roy Gould	Harvard Center for Astrophysics
Seth Tissue	Lead Developer of NetLogo
Mike Hansen	Middle School Teacher, Malden MA

2011-2012 Field Test teachers

Jenelle Hopkins	Centennial High School, Las Vegas NV
Jim Lindsey	Mooresville High School, Mooresville, IN
Rick Dees	Huntley Project High School, Worden, MT
Vic Hunt	Lenape Regional High School, Shamong, NJ
Beth Spear	Central High School, Salem WI
Peter Schwartz	Grey Culbreth Middle School, Chapel Hill, NC
Lacey Huffling	Arborbrook Christian High School, Mathews, NC
Andrea Williams	Orchard Lake Middle School, West Bloomfield, MI
Joshua Abernethy	Randolph Early College High School, Asheboro, NC
Leslie Knight	Framingham High school, Framingham MA
Sarah Tomkinson	Framingham High School, Framingham MA
*Jon Krawiec	Waterville Central High School, Waterville, NY
*Mark Case	Southern Guilford High School, Greensboro, NC
*Ruth-Joy Stephenson	P.S. 235 Lenox School, Brooklyn NY

2010-2011 Field Test teachers

Sarah Kehoe	Framingham High School, Framingham, MA
Leslie Knight	Framingham High School, Framingham, MA
Jon Krawiec	Waterville Central High School, Waterville, NY
Ruth-Joy Stephenson-John	P.S. 235 Lenox School, Brooklyn, NY
*Tobias Hatten	The Village School, Great Neck, NY
*Kathy Bertrand	Pierce Middle School, Milton, MA
*Jill Markarian	Pierce Middle School, Milton, MA
Jennifer Sundstrom	Ottoson Middle School, Arlington, MA
Heather Krepelka	Ottoson Middle School, Arlington, MA
Carol Feeney	Ottoson Middle School, Arlington, MA
*Jennifer Crafts	Ottoson Middle School, Arlington, MA
Travis Woodward	Ottoson Middle School, Arlington, MA
Brandon Bage	Ottoson Middle School, Arlington, MA

(* denotes teachers who did not fully participate)

Advisory Board:

Sarah Kehoe is a high school Earth Science teacher at Framingham High School, Framingham, MA.

Vanessa Bullard is a middle school Earth Science teacher at Belmont Middle School, Belmont, MA.

Marilyn Decker is the K-12 Science Director for the Milton Public Schools. She recently served as Director of Professional Development for Teachers 21. She was the K-12 Science Director for the Boston Public Schools from 2001-2008.

Marcia Linn is a professor of development and cognition, specializing in education in mathematics, science, and technology, in the Graduate School of Education at the University of California, Berkeley.

Dan Murray is a professor of research, Emeritus, Department of Geosciences, University of Rhode Island and Principal Investigator for the Rhode Island Technology Enhanced Science (RITES) program, a targeted NSF MSP project.

Ron Snell is a Professor of Astrophysics at the University of Massachusetts, Amherst; he uses radio astronomy in research on molecular clouds and star formation.

Contacts

Phoebe Cohen contacted us to set up the “Is there life in space?” investigation for her to use in an undergraduate astrobiology course at MIT.

Cornelia Harris, Celia Cuomo, and Alan Berkowitz, of the Cary Institute for Ecosystems Studies at Marist College used parts of the High-Adventure Science investigations. Dr. Harris and Dr. Cuomo used the "Will there be enough fresh water?" investigation and models from the Modeling Climate Change investigation in their undergraduate curriculum. Additionally, we have had conversations regarding future collaborative work focused on biodiversity and the High-Adventures Science framework.

Ten middle and high school teachers have contacted us and implemented High-Adventure Science in their curriculum.

Dr. Tamara Ledley, PI on the CLEAN project funded by NSF, participated in a CLEAN materials evaluation panel. I am a member of the CLEAN listserv and co-presented work at the DRK12 meeting.

1. Describe the major research and education activities of the project.

Summary of Project Activities

The goal of the High-Adventure Science exploratory DRK-12 project was to bring the excitement of frontier science into the classroom by allowing students to explore pressing unanswered questions in Earth and space science that scientists around the world are currently investigating. The High-Adventure Science (HAS) project has students investigate the mechanisms of climate change, learn how scientists use modern tools to find planets around distant stars, and evaluate whether underground stores of fresh water will be sufficient to support growing populations.

The High-Adventure Science project has created computer-based investigations around each of these topics. Each investigation is designed for five class periods and includes interactive computational models, real-world data, and a video of a scientist discussing his or her computer-based research on the same unanswered questions. While we did not expect the students to solve the problems posed in the curriculum, our goal was to have students experience doing science the way scientists do. It is the approach that mattered—one based on students critically thinking about evidence, making predictions, formulating explanations, drawing conclusions, and qualifying the level of certainty with their conclusions. The curriculum therefore focused on helping students make claims, defend their claims, and express their levels and sources of certainty with the claims. The research on the project focuses on measuring students' critical thinking by having the students formulate explanations and justifications to support their claims.

To accomplish these goals, the project activities included:

Developing Materials: The project produced three five-day investigations. The investigations include scaffolded computational models that enabled students to experiment with the Earth system under study through guided exploration of the models, real-world data related to the content, and videos of scientists who use models in their research of the topic.

Development and Validation of Assessment Items: To examine how students develop their critical thinking abilities when they make claims based on evidence, we developed and validated new explanation-certainty item sets. These item sets consist of four separate questions that require students to 1) make a scientific claim (claim), 2) explain their claim based on evidence (explanation), 3) express their level of certainty (certainty), and 4) describe the source of certainty (certainty rationale).

Current Event Blog: Because the High-Adventure Science program focuses on big unknown questions in science, we started a blog to show how our materials relate to what scientists are doing to answer these big questions. Science stories were pulled from science news websites.

Formative Testing and Revision: The materials were field-tested twice in diverse class settings. In year two of the project, nine teachers tested the climate change investigation. Some of the teachers also tested the search for life in space investigation. In year three, eleven teachers tested at least two of the investigations each. The materials were revised on the basis of the findings of the year two field tests.

Technology Development: The High-Adventure Science project has made extensive use of NetLogo to create the key interactive models. In addition, the Investigations Portal, previously developed by The Concord Consortium, was upgraded to support the functions needed to monitor and assess students' performance remotely.

Professional Development: The project held two-day summer workshops prior to each field test for participating teachers (Summer 2010 and Summer 2011). We provided extensive online support through a private Facebook group and individualized e-mails. All teachers participated on the HAS Teachers 2011-2012 Facebook group, posting information about their anticipated start and end dates for particular curricular units, as well as information that they thought the larger group might find helpful/interesting (links to related websites, in-class demonstrations and labs). Additionally, we provided teacher guides to help teachers with implementation.

Dissemination: The project actively disseminated the materials and research findings through presentations, newsletter articles, papers, and workshops.

Project Rationale

The goal of this project was to investigate a method of for injecting contemporary science into classrooms by engaging students in unanswered questions that scientists around the world are currently exploring.

Inspired by *Science*'s 125th year special issue, "125 Questions: What Don't We Know?" (July 2005), the purpose of this project was to explore whether it is possible to generate excitement and motivation in middle and high school students by giving them a taste of the unknowns in selected science topics, doing it in a way that students can understand and that is simultaneously engaging, inviting, and matches core standards.

The way that students learn about unsolved science topics needs to reflect the way science proceeds; students cannot actually perform the scientific experiments, but they can explore aspects of them by using computational models. Students can experiment with models and learn deeper concepts by exploring the emergent phenomena. The High-Adventure Science project is one part of ongoing research and development at The Concord Consortium to take full advantage of computer technologies for exploring science and to measure the impact of the intervention on students' thinking about the process of science.

Curriculum Development

The High-Adventure Science project created three investigations for middle and high school students that focus on current, compelling, unanswered questions in Earth and Space science:

- What will Earth's climate be in the future?
- Is there life in space?

- Will there be enough fresh water?

Included in each investigation are a video that highlights a scientist in the field, unique NetLogo-based computational models, and assessment tools focused on students' argumentation skills. The topics were selected based on based on teacher and student interest surveys, an analysis of curriculum balance, correlation to standards, and modeling capacity.

The Investigations

1. "Modeling Earth's Climate" Investigation

This investigation focused on the question: *What will Earth's climate be in the future?* In this investigation, students explore past climate changes and learn how mechanisms for positive and negative feedback can affect global temperature. They think about how scientists use this information to make climate change predictions. Students learn about where there is certainty in the climate data and where there is uncertainty with regard to predicting what will happen. This investigation pays special attention to helping students think about the presented evidence and how to evaluate the conclusions scientists can draw from the evidence.

Students explore data from NASA and the Vostok ice cores and look at trends over different time scales. They begin to explore the limitations of conclusions drawn from the data. Then students interact with models to learn about how radiation interacts with Earth's surface and atmosphere, the relationship between ocean surface temperature and carbon dioxide sequestration, the relationship between atmospheric carbon dioxide levels and the amount of water vapor, and, in the final model, the relationship between all three (carbon dioxide, ocean surface temperature, and water vapor). Additionally, students explore albedo, changing the amount of ice and cloud cover in their models to examine how different surfaces provide negative and positive feedbacks to the temperature increases resulting from increased levels of greenhouse gases. Finally, students explore how all the variables interact with each other to produce global temperature effects.

2. "Is there life in space?" Investigation

The second investigation focused on the question: *What is the probability of finding life outside of Earth?* The main focus of this unit is student exploration of planet-hunting methods using a dynamic model that simulates a single planet orbiting a star. The uncertainty questions focus on data interpretation and being able to detect faint to moderate signals in noisy data.

Students were introduced to the transit method and the radial velocity method of planet-hunting. The transit method involves interpreting light intensity data from a star in an attempt to observe a periodic drop in brightness. Students explore factors such as planet size, the angle of orbit with respect to the observer, and the precision of the light-sensing instrument on scientist's ability to detect planets via the transit method. Students are also introduced to the radial velocity, or wobble, method of detecting planets. This method involves interpreting the shift in the apparent wavelengths of light coming from a star; as the planet moves around the star, it exerts a gravitational pull, resulting in a star wobble. Students use models to explore the effects of planetary mass on a star's motion, changes in wavelengths of light as related to star motion, and how the angle of orbit influences a scientist's ability to detect a shift in the wavelength.

Finally, the investigation explores conditions for habitability. Students look at properties of five different star types and the zone of habitability around each star. Students end the investigations with a focus on how telescopes can be used to analyze light from a star to look at planetary atmospheres and how this information might reveal clues about which planets are more likely to be habitable.

3. "Will there be enough water?" Investigation

The third investigation focused on the question: *Will there be enough freshwater resources for Earth's growing population?* The main focus of this investigation is to have students explore Earth's freshwater resources: where they can be found, how we use them, and why we must think about sustainable use as Earth's population increases. The investigation ultimately explores why human and ecological needs should be balanced and how freshwater resource issues vary around the world.

Students begin by exploring parts of the water cycle: groundwater flow and recharge, evapotranspiration, and precipitation. With the model, students are able follow water through the water cycle. Students evaluate how the supply and demand for fresh water differs around the world. Students then explore the movement of water though the ground; models show how water moves through substances of different permeability.

Students use models to explore how aquifers are created. The models enable students to investigate how the level of the water table affects the water level in streams and ponds. Students experiment with creating different subsurface layer configurations to look at the formation of water tables and aquifers. Finally, students focus on the relationship between groundwater recharge, related to permeability and porosity, and the rate at which water is pumped out for human use. Students are introduced to some ways in which humans have disrupted the water cycle and are challenged to suggest solutions to a freshwater availability problem.

These investigations can be seen by clicking on the Project Portal link at:
http://www.concord.org/projects/high-adventure-science

Teacher guides for the investigations can be found at:
http://www.concord.org/projects/high-adventure-science#participants

Year One Development

The first year was devoted primarily to narrowing down the topics for the investigation, developing the computational models, and creating the first drafts of the investigations. We also recruited teachers and developed and validated the assessment items.

Our approach was to sketch out a large number of possible topics and then winnow them down based on input from teachers, students, and content experts. We presented possible topics to five classes of ninth grade students, held a model design brainstorming session with developers at The Concord Consortium, correlated the topics to national science standards, and held a focus group with Earth science teachers from three local Massachusetts school districts. The three topics that were developed came out as a result of these efforts.

We developed outlines for each of the three investigations and created the interactive model for the Modeling Earth's Climate investigation. A draft of the five-day curriculum was completed for use in the first summer teacher workshop.

A great deal of work was done to delineate what constitutes acceptable evidence of students' achievement of the desired results. Dr. Lee developed the explanation-certainty item sets and scoring rubrics that measure students' understanding of Earth science concepts in the context of frontier science and students' argumentation skills, including students' ability to deal with uncertainty in science.

These items sets were designed to reveal a more complete picture of student understanding. Following a scientific claim, students must answer a question and explain their reasoning. Students' explanations help us understand how they think about both the evidence and the claim. Certainty rationale items measure whether or not students recognize the source of uncertainty of their claims. Through repeated exposure, our goal was to encourage students to reflect on both the evidence that they generated from using the models and the real-world data and to evaluate how certain they are about their own claims, as well as the claims of scientists. The item sets were piloted in May 2010. Results from our pilot indicated that students who could make multiple claims were likely to consider evidence from the models and data from the scientists, Additionally, students uncertainty in scientific argumentation transitioned from self-concepts (they personally were uncertain) to scientific uncertainty (when data was inconclusive).

In the first year, we recruited teachers to participate in the High-Adventure Science project by posting to several listservs targeting Earth Science teachers, including ESPRIT and MESA (Massachusetts Earth Science Alliance).

Year Two Development

In year two, the remaining two investigations, "Will there be enough fresh water?" and "Is there life in space?" were completed and field-tested along with the "Modeling Earth's Climate" investigation. This involved developing uniquely complex NetLogo models that required new capacity from the software.

Additionally, we filmed videos of scientists for the investigations. For the "Modeling Earth's Climate" investigation, we created a video entitled "Climate Modeling: Using History to Inform the Future." This video features Dr. Mark Chandler, a climate scientist at the National Aeronautics and Space Administration (NASA) Goddard Space Center in New York. For the "Will there be Enough Fresh Water?" investigation, we filmed a video entitled "Using Water Responsibly." We interviewed Dr. Holly Michael, a groundwater hydrogeologist at the University of Delaware. The "Is there life in space?" investigation includes a video, used with permission, from the NOVA ScienceNOW group at the Corporation for Public Broadcasting.

To recruit more teachers for the second implementation year of the project, we contacted the Executive Director of the National Earth Science Teachers Association (NESTA), Dr. Roberta Johnson Killeen. We also contacted the Executive Director of the National Association of Geoscience Teachers (NAGT), Cathryn Manduca. Additionally, a notice about the High-Adventure Science project was posted on the National Earth Science Teachers Association's Facebook page. (http://www.facebook.com/group.php?gid=40697591874&v=wall).

Finally, the High-Adventure Science project expanded and extended the functionality of our web-based portal as described in the technology section that follows.

Year Three Development

Prior to the implementation of the High-Adventure Science investigations in classrooms, the staff substantially revised the materials and assessments based on feedback and results from year two field tests. Changes in the investigations were in the following categories:

Readability

Mike Hansen, a sixth-grade Earth science teacher in Malden, MA, reviewed each of the investigations. Mr. Hansen sat with Concord Consortium staff and read through all three investigations, giving feedback on tone, readability, and accessibility of the materials to students. Mr. Hansen's feedback, analysis of students' responses to questions from year two field-testing, and teachers' feedback from year two field-testing were used to revise the text and models in all three curricular investigations.

Inclusion of more Explanation-Certainty Item Sets

Explanation-certainty item sets proved to be quite informative about student thinking regarding content and process skills. However, it seemed that the explanation-certainty items in the pilot versions focused only on topics on which there was a low amount of certainty with data or models. In order to give students a wider range of experiences, additional item sets were added to the curriculum around topics for which students could have greater certainty (more complete data sets). For example in the "Will there be enough water?" investigations students interact with a model of the water cycle, students are then asked "When water is absorbed into the ground, is it trapped in the ground?" Students can use evidence from the model and their exploring the path of water to answer the question and explain their certainty of the answer. This is different then the more open-ended question "Sustainable water use occurs when the withdrawals of water are equal to the inputs of water, which pumps in the model show sustainable water use?" Students must rely on experimentation and evidence from the model to explain their certainty with their answer.

In the “Will there be enough fresh water?” investigation, explanation-certainty item sets focused on the prediction of water flow in sediments of differing permeability and porosity and on the relationship between human water use, sediment structure, and precipitation.

In the “Is there life in space?” investigation, explanation-certainty item sets were geared towards measuring students’ understanding planet-hunting methods, as well as interpretation of data from telescopes and the probability of finding life outside of Earth.

Likewise, explanation-certainty item-sets in the “Modeling Climate Change” investigation were aimed at helping students to focus on the positive and negative feedback loops modeled in the curriculum and managing the inherent uncertainty that comes with trying to predict future events.

Model/curriculum modification

In addition to the assessment and readability modifications, we revised the models within each investigation and added videos of scientists to the curricular units.

For all three investigations, we made many different versions of models. Models early in each investigation involve simple interactions; later models introduce more complexity as students gain content knowledge and are better able to interpret the results from more complex models. The ending models in each investigation, therefore, are the most complex, both in representation and in interaction.

Curriculum modifications for the “Will there be enough fresh water?” investigation changed as a result of reviews and field-testing. Changes included increased emphasis on the water cycle and how water moves through the ground and models and visualizations focused on permeability and porosity of the sediments. These topics correspond directly to concepts in the traditional Earth science curriculum.

The space investigation revisions included new content covering spectroscopic analysis of planetary atmospheres, a brief discussion of what elements/compounds might indicate about the presence of life or possibility of life, a starting emphasis on what is currently known about planet hunting, and a focus on habitability and the Goldilocks effect.

Scientific Review

All of the assessments, curriculum materials, and teacher guides were reviewed by scientists for scientific accuracy and pedagogical validity. Their feedback was used to revise the investigations before the field tests in year three. Dr. Mark Chandler, a scientist at the NASA Goddard Space Center, reviewed the “Modeling Climate Change” investigation. Dr. Roy Gould, a scientist at the Harvard Center for Astrophysics, reviewed the “Is there life in space?” investigation. Dr. Holly Michael, a scientist at the University of Delaware, reviewed the “Will there be enough fresh water?” investigation.

High-Adventure Science Blog

Because the High-Adventure Science program focuses on big unknown questions in science, we started a blog to show how our materials relate to what scientists are doing to answer these big questions. Science stories were pulled from science news websites and popular media. We encouraged the High-Adventure Science teachers to assign blogs to their students for reading and posting comments. We have no evidence from blog postings that students read these blog posts. However, since the blog posts were also posted to the HAS Teachers Facebook group, we do have evidence that teachers read the postings from their comments on the Facebook group.

Table 1: Links to High-Adventure Science Blog Posts

Blog Post	**URL**
Climate Change: Back to the Future?	http://blog.concord.org/climate-change-back-to-the-future
Surprising effects of solar activity on Earth's temperature	http://blog.concord.org/surprising-effects-of-solar-activity-on-earths-temperature
Climate and Pollution	http://blog.concord.org/climate-and-pollution
Certainty	http://blog.concord.org/certainty
Goldilocks and the Habitable Planets?	http://blog.concord.org/goldilocks-and-the-habitable-planets
Carbon dioxide as a structural component?	http://blog.concord.org/carbon-dioxide-as-a-structural-component
Tracking the Permafrost Line	http://blog.concord.org/tracking-the-permafrost-line
Science and Politics: What to do?	http://blog.concord.org/tracking-the-permafrost-line
Science and Politics: What to do?	http://blog.concord.org/science-and-politics-what-to-do
Burning the rainforest to cool the globe	http://blog.concord.org/burning-the-rainforest-to-cool-the-globe
How much does a star weigh?	http://blog.concord.org/how-much-does-a-star-
Missing: Fresh Groundwater	http://blog.concord.org/missing-fresh-groundwater
Finding a needle in a haystack, how to deal with noise in the data	http://blog.concord.org/finding-a-needle-in-a-haystack-how-to-deal-with-noise-in-the-data
How can you tell what's in the atmosphere of a planet that's over one billion miles from Earth?	http://blog.concord.org/how-can-you-tell-whats-in-the-atmosphere-of-a-planet-thats-over-one-billion-miles-from-earth
Finding little planets with new technology	http://blog.concord.org/finding-little-planets-with-new-technology

Know thy star to know its planets	http://blog.concord.org/know-thy-star-to-know-its-planets
It’s going to be a warm one in the south	http://blog.concord.org/its-going-to-be-a-warm-one-in-the-south
Slow down glacial flow with warmer summers?	http://blog.concord.org/slow-down-glacial-flow-with-warmer-summers
Finding other ,”Earths”	http://blog.concord.org/finding-other-earths
Going up?	http://blog.concord.org/going-up
The frozen tundra could heat the Earth	http://blog.concord.org/the-frozen-tundra-could-heat-the-earth
Reading layers when layers are disturbed	http://blog.concord.org/reading-layers-when-layers-are-disturbed
Thinking like a scientist	http://blog.concord.org/thinking-like-a-scientist
Poison helping to develop life?	http://blog.concord.org/poison-helping-to-develop-life
Trees to the (partial) rescue!	http://blog.concord.org/trees-to-the-partial-rescue
Wanted: Cause of the End of ”Snowball Earth”	http://blog.concord.org/wanted-cause-of-the-end-of-snowball-earth
Ocean Currents’ The Big Unknowns	http://blog.concord.org/ocean-currents-the-big-unknowns
A Red ”Snow White”	http://blog.concord.org/a-red-snow-white
Causality: How to Interpret Graphs	http://blog.concord.org/causality-how-to-interpret-graphs
What makes scientists more certain?	http://blog.concord.org/what-makes-scientists-more-certain
Raising the water table the natural way	http://blog.concord.org/raising-the-water-table-the-natural-way
Digging into Permafrost	http://blog.concord.org/digging-into-permafrost
Harvesting Planets	http://blog.concord.org/harvesting-planets
Good Science/Bad Science	http://blog.concord.org/good-sciencebad-science
Irrigation and Climate Change	http://blog.concord.org/irrigation-and-climate-change

Pumice: Islands of Life?	http://blog.concord.org/pumice-islands-of-life
Transpire Locally, Cool Globally	http://blog.concord.org/transpire-locally-cool-globally
Finding Fossil Aquifers on Earth	http://blog.concord.org/finding-fossil-aquifers-on-earth
Absolute Certainty Is Not Scientific	http://blog.concord.org/absolute-certainty-is-not-scientific
When More Is More	http://blog.concord.org/when-more-is-more
What caused the Paleocene-Eocene Thermal Maximum?	http://blog.concord.org/what-caused-the-paleocene-eocene-thermal-maximum
More planets!	http://blog.concord.org/more-planets
When in Drought!¶	http://blog.concord.org/when-in-drought
The Great Antarctic Glaciation	http://blog.concord.org/the-great-antarctic-glaciation
Using Dynamic Models to Discover the Past (and the Future?)	http://blog.concord.org/using-dynamic-models-to-discover-the-past-and-the-future

Field testing

School demographic distribution

Table 2 below describes the diversity of school settings in which the High-Adventure Science curriculum was field-tested. The schools represent a wide distribution of locations, student demographics, and grade levels.

Table 2: Demographic Information (as of 2009-2010 school year)

School	# of students	American Indian /Alaskan	Asian/ Pacific Island	Black	Hispanic	White	Two or More Races	Free & Reduced Lunch
Centennial High School, Las Vegas NV	2935	0.6%	7.4%	15.3%	19.4%	57.4%	0.0%	19.4%
Mooresville High School, Mooresville, IN	1355	0.4%	0.5%	0.2%	1.1%	96.7%	0.0%	25.6%
Huntley Project High School, Worden, MT	240	3.3%	1.3%	1.3%	4.6%	89.6%	0.0%	27.1%
Lenape Regional High School, Shamong, NJ	850	0.4%	3.2%	1.5%	6.9%	88.0%	0.0%	7.2%
Central High School, Salem WI	1201	0.6%	1.2%	1.5%	4.0%	92.7%	0.0%	17.7%
Grey Culbreth Middle School, Chapel Hill, NC	645	0.3%	8.2%	16.0%	6.5%	69.0%	0.0%	18.8%
Arborbrook Christian High School, Mathews, NC		NA	NA	NA	NA	NA	NA	NA
Orchard Lake Middle School, West Bloomfield, MI	782	0.0%	15.3%	34.3%	0.4%	48.3%	0.0%	21.5%
Randolph Early College High School, Asheboro, NC	319	0.3%	1.6%	7.8%	16.6%	73.7%	0.0%	33.5%
Framingham High school, Framingham MA	2190	0.3%	5.9%	8.5%	17.7%	67.3%	0.3%	27.1%
Waterville Central High School, Waterville, NY	427	0.5%	0.0%	2.3%	0.0%	97.2%	0.0%	36.3%
Southern Guilford High School, Greensboro, NC	1014	0.8%	7.0%	47.1%	8.9%	36.2%	0.0%	56.6%
P.S. 235 Lenox School, Brooklyn NY	1383	0.1%	1.4%	94.7%	3.0%	0.7%	0.0%	NA
The Village School, Great Neck, NY	44	0.0%	2.3%	4.5%	6.8%	86.4%	0.0%	2.3%

Pierce Middle School, Milton, MA	860	0.1%	4.2%	20.9%	3.6%	69.1%	2.1%	15.9%
Ottoson Middle School, Arlington MA	1060	0.1%	7.6 %	3.8%	5.5%	80.9%	2.1 %	11.4%

Year two field-testing

Our field-test teachers were asked to test one or two of the investigations, administering a pre-test and a nature of science survey at the beginning of the year, the curricular unit(s), and separate pre-and post-tests for each of the investigations.

The teachers attended a professional development workshop in which we gave teachers our expectations; teachers explored the investigations and participated in discussions regarding how to teach about unanswered scientific questions and uncertainty.

We helped teachers learn how to set up classes for collecting and managing students' data online. We trained 13 teachers; nine of the teachers became active field test teachers. Each of the nine teachers field-tested the climate change investigation in year two.

Results from the field test influenced changes in the curriculum content, as previously described. We also revised assessment items to more closely match the curriculum. Additionally, we planned a change in the length of the teacher professional development, to give teachers more time to focus on the nature of science content of the High-Adventure Science curriculum.

Year three field-testing

Our field-test teachers were asked to test two or three of the investigations, administering a pretest, with questions covering content related to all three investigations, and a nature of science survey at the beginning of the year, the curricular units, a separate pre-and post-test for each of the investigations, and an end-of-the-year post-test (with the same questions as the beginning-of-the-year pre-test). The Table 3 below indicates the investigations completed by the participating year three field-test teachers.

Teacher	School	Climate	Water	Space
Jenelle Hopkins	Centennial High School, NV	X	X	X
Jim Lindsey	Mooresville High School, IN	X	X	X
Rick Dees	Huntley Project High School, MT		X	X
Vic Hunt	Lenape Regional High School, NJ		X	X
Beth Spear	Central High School, WI	X		X
Peter Schwartz	Grey Culbreth Middle School, NC	X		
Lacey Huffling	Arborbrook Christian High School, NC	X	X	X
Andrea Williams	Orchard Lake Middle School, MI	X	X	X
Joshua Abernethy	Randolph Early College High School, NC	X	X	
Leslie Knight	Framingham High School, MA	X	X	
Sarah Tomkinson	Framingham High School, MA	X	X	
Jon Krawiec	Waterville Central High School NY			X

Table 3: Investigations completed by year three teachers.

The teachers attended a two-day workshop, held on August 1-2, 2011 in Concord, MA. The foci of the workshop included the following:

- Explore High-Adventure Science curriculum.
- Develop teaching strategies for using materials.
- Support the research for the project.
- Be prepared for the school implementation.
- Develop a community.

During the workshop, teachers had an opportunity to work through each of the curriculum units in detail. We held sessions on teaching strategies specifically focused on teaching with computational models, getting students to question the data and outcomes of the models, teaching about the unknown, the explanation-certainty item sets and how they can reveal student thinking, and using the High-Adventure Science Facebook group to develop community support. In addition, we helped teachers set up classes and learn how to access their students' work. We discussed strategies for grading, differentiation of instruction, and expectations for being a field-test teacher and giving feedback.

Each teacher was asked to administer the beginning-of-the-year pre-test and the nature of science survey as early as possible in the school year. Teachers committed to implementing the investigations as their curricular schedule allowed. This meant that the investigations happened at different times during the year. Teachers implemented individual pre- and post-tests for each of the curriculum investigations.

The table above indicates the investigations that each teacher implemented. At the end of the year, Jon Krawiec was unable to complete his intended commitment to the project; Mr. Krawiec ran into issues regarding scheduling time in the computer lab. He was unable to complete the High-Adventure Science project this year. Similarly, Peter Schwartz was unable to complete his intended commitment.

Technology Development

Year one

The High-Adventure Science project developed a website for sharing the project's development with teachers, researchers, and the general public. The website http://www.concord.org/projects/high-adventure-science has several pages.

Home page: The home page describes the project and provides a link to the portal where a user can preview the activities or sign up to use them, a link to the blog of science news stories related to the curriculum investigations, and links specifically for teachers and researchers to get more in-depth information about the project and its research.

Research page: This page includes an overview of our research questions and a description of our research tools.

For Teachers page: This page includes teacher guides for the investigations, links to the investigations, and information about the technology requirements.

Publications and Videos page: This page includes links to the videos created for the investigations and links to the papers written for this project.

Year two

The High Adventure Science project made extensive use of NetLogo to create the key interactive models used for modeling planet hunting, groundwater hydrology, and climate variable interactions.

The agent-based programming and simplicity of the language meant that programming in NetLogo did not require professional programmers; as an exploratory project, this was key to being able to create valuable models quickly. Initially, we intended to use NetLogo as a prototyping environment that would allow us to quickly try out different ideas before committing to production coding in another language. We quickly discovered that NetLogo prototypes, with some polishing, could be used in the final activities of the project. It also meant that educational content experts could produce the code. As a result, the models developed were innovative, scientifically correct, and educationally sound. Three members of the High-Adventure Science team became expert NetLogo developers and developed all the NetLogo models used in High-Adventure Science. This greatly speeded the development, testing, and integration of models and increased the model functionality far beyond what we had initially thought possible.

Our use of NetLogo for complex models was unusual and led to huge programs that were probably among the largest NetLogo programs then in existence. We got NetLogo to add some unique capacities. The planet-hunting model required additional functionality, including more 3D capacity, the ability to view inserted more-detailed views in the model, and “soft keys” that are managed under programmatic controls. The soft keys were important ways to simplify the user interface and simplify the interactions for student learners. Our use of NetLogo as a development language has pushed the limits of the language, but fortunately we have had excellent collaboration with the team from Northwestern University, under Uri Wilensky, that developed and continues to support NetLogo.

During year two, the High-Adventure Science Project extended and expanded the functionality of The Concord Consortium’s Investigations-based web-based portal. The High-Adventure Science project focused on improving the user interface and making detailed reports available for teachers and researchers. This was done by developing a way for teachers to customize the student work they want to see for assessment and developing methods for displaying student work online. Improvements included automatic scoring of multiple choice items, a “Cover Flow”-like view allowing for quick perusal of student-generated images, and an organized view of open-ended responses.

In addition, the High-Adventure Science project developed a way to generate researcher reports. These reports collate data across teachers, classes, and schools for each of the investigations. The data is exported to files that are easily imported into spreadsheets for scoring and statistical analysis.

Year three

Year three saw minimal additions to the technology as this year was focused on field-testing the revised activities and conducting research. Additions in year three are as follows:

- Updating the portal to make it easier for public and private school teachers to register
- Fixing the jnlp launching system
- Fixing several issues that caused data loss due to multiple launching of activities
- Improving download capacity for extra-large amounts of data
- Revising methods for filtering data using date-ranges, schools, and activities (necessary for creating researcher reports)

Technology Support for Schools

Technology support generally fell into two categories: firewall issues and software bugs. We worked with schools to make sure that the proper software was installed, that their computers could reach our servers, and to eliminate bugs in the modeling and data collection software. We communicated directly with technology coordinators in schools and with teachers themselves. A great deal of energy was put into recovering student work, lost passwords, and issues regarding non field-test teacher registration issues.

Dissemination

The project actively disseminated the materials and research and findings through presentations, newsletter articles, workshops and meetings. The project website (http://www.concord.org/projects/high-adventure-science) was updated throughout the project's lifespan to include the newest investigations, blogs, and articles.

High-Adventure Science has been used in several schools and universities by teachers not part of the field-test. The Investigations were used in freshman courses at MIT and Marist College. Ten middle and high school teachers implemented the investigations in their curriculum after hearing about the material in workshops and from the Science Teacher publications. Additionally, approximately 30 teachers explored the investigations as part of teacher professional development workshops delievered by other NSF sponsored projects.

We published articles in The Concord Consortium biannual @Concord newsletter, which is distributed in print form to 8,000 teachers and administrators and is available online on the Concord Consortium website. Additionally, two articles were written for *The Science Teacher*, and the work was presented at several conferences. Below is a listing of papers and presentations resulting from this project:

Publications

Pallant, A, Lee, H-S, & Pryputniewicz, S. (2012). Systems Thinking and Modeling and Climate Change. Accepted by *The Science Teacher,* to be published in October 2012.

Pallant, A, Lee, H-S, & Pryputniewicz, S. (2012). Exploring the Unknown. *The Science Teacher.* Vol 79, No. 3.

Pallant, A. (2011). Looking at the evidence. What we know. How certain are we *@Concord 15*(1), 4-6.

Pallant, A. (2010). Modeling the unknown is high adventure. *@Concord*. 14 (1), 6-7. P

Presentations

*Mapping DRK12 Project Activities to Climate and Environmental Literacy Principles.*DRK12 PI Meeting, Washington DC, June 13-15, 2012

Uncertain Answers: Exploring Climate Change and Water Sustainability with Models. National Science Teachers Association, Indianapolis, IN, March 30, 2012.

Looking at the Evidence: How certain are we? American Association for The Advancement of Science (AAAS), Vancouver, BC, February 17, 2012.

Interactive Models for Exploring Planet Discovery and Extraterrestrial Life, Space Exploration Educators Conference (SEEC), Houston, TX, February 1-3, 2012

Complexity of Modeling. Santa Fe Institute Summer program for high school students. July 15, 2011

Pallant, A., & Lee, H-S. (2011, April) Characterizing uncertainty associated with middle school students' scientific arguments. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching (NARST), Orlando, FL.

Online STEM Initiatives: A Hands-On Training Workshop, Virtual School Symposium, Glendale, AZ, November 14, 2010.

Inquiry in the Digital Age, Enhancing Science Learning using Computer Models, Cyberlearning Tools for STEM Education, Berkeley, CA, March 9, 2011.

Online Courses and Materials That Provide True Technology Integration Across the Sciences, National Science Teachers Association, San Francisco, CA, March 10, 2011.

Linking Student Achievement, Teacher Professional Development, and the Use of Inquiry-based Computer Models in Science, National Science Teachers Association, San Francisco, CA, March 12, 2011.

Project Evaluation and Advisory Board

The project's advisory committee met on June 4, 2010 and on May 26-27, 2011 to review the project's progress and to examine the annual National Science Foundation reports from The Concord Consortium. All the advisors listed above attended the meetings. In addition to being evaluators, the advisory board members also played the role of external evaluators.

During the initial 2010 meeting, the advisors were given an overview of the project goals and the work we had been doing. They reviewed the "Modeling Earth' Climate" prototype investigation and learned about the portal system for collecting data and the research and field-test plan. The following describes key suggestions they made to the project:

- Stay focused on the learning goals. The investigations contain more information than could be taught thoroughly in a 5-day module.
- Make connections to students' prior knowledge and experience.
- Establish activities that link the nature of science to the disciplinary knowledge.
- Illustrate measurement errors and other aspects of uncertainty.
- Highlight the uncertainty that exists for the topic area so students can get a clear understanding of uncertainty. What do we know about and what do we not know about?
- Make sure the questions really ask for using evidence to explain reasoning.

During the second meeting, the advisors were again given a progress report. They explored all three investigations and gave feedback on content and pedagogy. They learned about research results from the first field-test and discussed follow-on proposal ideas and papers.

During the second meeting, board members said that they felt that the project had made good progress towards its objectives and found that the project developed the necessary tools to measure the impact of the curriculum units on student learning. They liked the progress the project had made on focusing on science as a process, and they were interested in how students learn about drawing conclusions from the data and evidence presented to them. Board members did express concern about the scope of the modules and the content load. Additionally, they provided recommendations for video subjects and ways to make the curriculum more successful. The board saw much potential for this project expanding beyond the three curriculum units in both Earth science domains as well as in other subject domains.

Educational Research Activities

Year one. Uncertainty and Scientific Argumentation Assessment Design and Testing

Research Questions

Scientific argumentation consists of claim and justification and can happen in either rhetorical or dialogic form. Toulmin (1958) specified that an argument may include up to six elements such as claim, data, warrant, backing, qualifier, and rebuttal. Toulmin's specification has resulted in various analytic methods that examined students' arguments expressed in written artifacts as well as online, small-group, or classroom discourse patterns (Sampson & Clark, 2008). The most analyzed elements of Toulmin's argument structure have been claim, data, warrant, and backing. Rebuttals were occasionally studied in dialogic discourse where one party detects weaknesses of the other party's argument. The presence of rebuttals was considered as evidence for a higher level of students' scientific argumentation ability (Kuhn, 2010). On the other hand, the role that qualifiers play in students' construction of scientific arguments has attracted little attention. The qualifier in an argument modifies the degree of its certainty. Considering all scientific arguments involve uncertainty due to incomplete or insensitive measurements, limitations in current theory or model, and phenomena under investigation (AAAS, 1993), it is important to study the uncertainty associated with students' scientific arguments. In our year one study, we investigated:

- What types of uncertainty do students exhibit when formulating a scientific argument involving complex data sets typical in current science?
- How are students' uncertainty rating and rationale related to their knowledge and ability to coordinate claim with evidence?

Research Design

Theoretical framework. Toulmin (1958)'s argument structure provided the basis to design our assessment items for students' argumentation ability. Toulmin (1958) identified the following six elements (also see Figure 1):

- Claim (C) or conclusion "whose merits we are seeking to establish"
- Data (D) are "the facts we appeal to as a foundation for the claim"
- Warrants (W) "show that, taking these data as a starting point, the step to the original claim or conclusion is an appropriate and legitimate one"
- Modal qualifiers (Q) indicate "the strength conferred by the warrant" with

adverbs such as ‘necessarily’ ‘probably’ and ‘presumably’.

- Conditions of rebuttal (R) indicate “circumstances in which the general authority of the warrant would have to be set aside…exceptional conditions which might be capable of defeating or rebutting the warranted conclusion.”
- Backing (B) shows “assurances without which the warrants themselves would possess neither authority nor currency.”

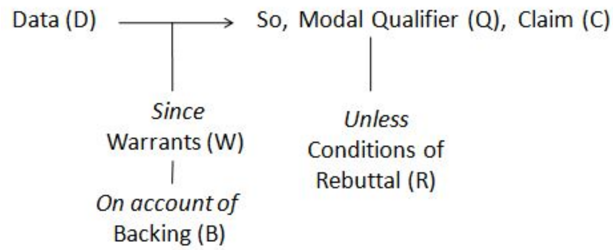


Figure 1. Toulmin’s argument structure (Toulmin, 1958, p.104).

Based on Toulmin’s argument structure (1958), we conceptualized the scientific argumentation construct consisting of six distinct levels (Table 4). We developed a construct map for students’ overall scientific argumentation ability based on claim, justification (data, warrants, and backing combined), uncertainty as modal qualifier, and conditions of rebuttal. Table 4 shows construct levels on a continuum in the order of increasing sophistication. Higher levels are assigned to students who include more elements in their scientific arguments. Using this construct map (Wilson, 2005), we hypothesized that items requiring the selection of a scientific claim (i.e., multiple-choice items) would be easier for students to answer than those requiring the elaborate coordination between claim and evidence (i.e., open-ended explanation items) and those requiring a scientific basis for explaining uncertainty involved in scientific arguments (i.e., uncertainty rationale items).

Table 4. A Construct Map for Scientific Argumentation Involving Claim, Explanation, Uncertainty Qualifier, and Conditions of Rebuttal.

	Description of the level	Toulmin (1958)	Student characteristics	Item design in this study
Level 1	Non-scientific			
Level 2	Scientific claim	Claim	Students think scientific claims can be made without support of evidence.	Claim
Level 3	Coordination between claim and evidence	Claim + data	Students recognize that adequate evidence is needed to support a claim.	Justification
Level 4	Reasoned coordination between claim and evidence	Claim + data + warrant/backing	Students can use theory or established knowledge to coordinate claim and evidence.	
Level 5	Modified, reasoned coordination between claim and evidence	Claim + data +warrant/backing + qualifier	Students recognize the uncertainty of claim given the strength of warrants.	Uncertainty
Level 6	Conditional, modified, reasoned coordination between claim and evidence	Claim + data +warrant/backing + qualifier + conditions of rebuttal	Students recognize conditions that the current claim may not be held by analyzing limitations related to measurements, current theory or model, and phenomena under investigation.	Conditions of rebuttal

Instrument Design. In developing items to measure students on the scientific argumentation construct, we selected items that correlated to the learning goals of the High-Adventure Science investigations. A total of 60 items were selected, modified from existing item sources, or newly designed: 50 multiple-choice items (16 climate change, 16 search for life in space, and 18 water resource), 5 explanation items (2, 2, 1 respectively), and 5 uncertainty rationale items (2, 2, 1, respectively). The content of the items was aligned with the national and state curriculum standards. The multiple-choice claim items were designed to measure students' overall knowledge, i.e. up to the level 1 on the scientific argument construct (Table 4), while the explanation items were designed and scored to measure up to the Level 3. The uncertainty rationale items were scored to measure the construct level 4. In the scientific argumentation item set, students were asked to claim, justify their claim, rate their uncertainty level on a five point Likert rating scale, and explain their uncertainty. See Figure 2.

Figure2. A scientific argumentation item set (Italics were added to indicate the item composition). The item was modified from TIMSS (IEA, 1995).

Jane and Mario were discussing what it might be like to live on other planets. Their science teacher gave them data about the Earth and an imaginary planet, Athena. The table shows these data.

[Claim]
Can life similar to life on Earth exist on Athena? (Select one) ___ Yes ___ No

	Earth	Athena
Atmospheric Conditions	21% oxygen	10% oxygen
	0.03% carbon dioxide	80% carbon dioxide
	78% nitrogen	5% nitrogen
	ozone layer	no ozone layer
Distance from a Star Like the Sun	148,640,000 km	103,600,000 km
Rotation on Axis	1 day	200 days
Revolution Around Sun	365 ¼ days	200 days

[Explanation]
Explain what might influence whether or not life can exist on Athena.

[Uncertainty]
How certain is your answer for life on Athena?

	Not at all certain (1)	(2)	(3)	(4)	Very certain (5)
	o	o	o	o	o

[Uncertainty Rationale]
Explain what factors influenced your uncertainty.

Data Collection and Analysis

We developed two test versions, each of which contained 25 multiple-choice items and 3 explanation items with 3 uncertainty rationale items. One explanation-uncertainty item set in Figure 2 appeared in both test versions. The tests were administered online to a total of 204 sixth-grade students: 129 of them taking Test A and 75 taking Test B. These students were sampled from 10 schools in the United States. We eliminated students who answered fewer than the first 15 items in the test, assuming they did not have enough time to answer the entire test. The resulting data set consisted of 120 students for Test A and 67 students for Test B. We scored student responses as follows:

- Multiple-choice claim items: (1) congruent with current scientific claim; (0) incongruent
- Open-ended justification items (See Figure 3):
  - (4) two or more theory-justified links between evidence and claim
  - (3) one theory-justified link between evidence and claim
  - (2) relevant pieces of evidence without theory-based justifications
  - (1) irrelevant pieces of evidence, scientifically-incorrect justifications, and non-normative ideas
  - (0) off-task/ blank
- Certainty rating items on a 5-point Likert scale
  - (2) certain: 4 or 5 rating
  - (1) 50-50: 3 rating
  - (0) uncertain: 1 or 2 rating
- Open-ended uncertainty rationale items (See Table 5):
  - (3) scientific uncertainty beyond investigation
  - (2) scientific uncertainty within investigation
  - (1) personal uncertainty

- (0) no information

Since we assumed a single construct, we conducted a Rasch analysis based on the Partial Credit Model, using the *Winstep* software (Linacre, 2010).

Figure 3. Explanation coding rubric

Relevant ideas related to evidence on the Life item shown on Figure 2:

- CO_2 idea (C): Athena has much more CO_2 than Earth.
- Oxygen idea (O): Athena has less oxygen than Earth.
- Rev-Rot idea (R): Rotation/Revolution comparison (Athena's revolution and rotation periods are the same).
- Ozone idea (OZ): Athena does not have an ozone layer.

Links (explains why each evidence piece is important using established scientific knowledge):

- C link: More CO_2 on Athena means hotter surface temperatures than Earth.
- R link: Athena's rotation and revolution periods are the same, indicating one side of the planet is always facing the sun and therefore is hot while the other side is always dark and cold.
- OZ link: Harmful UV rays are blocked by the ozone layer.

Explanation rubric	Criteria	Examples
Irrelevant (Score 0): Off-task	Blank OR Wrote some text unrelated to the item.	• Blank answers • Because I think so.
No-link (Score 1): Incorrect evidence	Elicited non-normative ideas or restated the claim selected.	• Nothing matches Earth. • It looks normal. • The details of Athena are relatively close to the details of EARTH.
Partial-link (Score 2): Relevant evidence	Elicited one or more ideas listed above.	• There is not enough oxygen and too much CO_2 . • There is too much carbon and too little oxygen and there is no ozone layer.
Full-link (Score 3): Single warrant between claim and evidence	Mentioned one of the links listed above.	• There is no ozone layer which means if life was to form it would most likely get burnt up by the stars radiation.
Complex-link (Score 4): Two or more warrants between claim and evidence	Mentioned two or more links mentioned above.	• The increased level of carbon dioxide would increase the greenhouse effect, and it is much closer to the sun than Earth, so it would be much hotter, like Venus, and so life could not live there. The lack of an ozone layer would also severely hurt life due to harmful UV rays reaching the surface of the Athena.

Table 5. Uncertainty rationale coding

Code	Description of coding categories
Minimal Information (Score 0)	• Blank for uncertainty rationale but answered the linked explanation item.
	• Wrote generic "I do not know" or similar answers • Provided off-task answers
	• Restated the scientific claim for the linked explanation item • Restated his/her uncertainty rating.
Personal Reasons (Score 1)	• Did not understand the question.
	• Did not possess general knowledge or ability for the question. • Did not learn/practice.
	• Did not know or had limited knowledge or understanding on particular scientific knowledge needed in the question.
	• Did not make sense of data/models presented.
Scientific Reasons (Score 2)	• Referred to "data/table" without mentioning specifics.
	• Referred to a particular piece of scientific knowledge or data.
	• Recognized the limitation of data/model provided in the question and suggested a need for additional data.
	• Stated that the scientific phenomenon addressed in the question is uncertain
	• Mentioned that current science such as models/knowledge/data about the scientific phenomenon addressed in the question are limited.

Year 2: Scientific Argumentation Validation Study and HAS Curriculum Study

Research Questions

In Year 1, we designed tests consisting of conventional multiple choice items and six scientific argumentation item sets. Results of the Year 1 assessment study produced promising results that the scientific argumentation item sets could cover the wider range of the scientific argumentation construct. Therefore, in Year 2, we designed instruments that exclusively consisted of claim, explanation, certainty, and certainty rationale items and conducted a scientific argumentation assessment validation study with a larger number of students (N = 956, compared to N = 203 in Year 1). Note that in the Year 1 assessment study, we did not psychometrically evaluate the uncertainty rating as part of the construct. After assessment validation, we used the same item designs to examine whether and how much students changed their performances on scientific argumentation item sets before and after the three HAS investigations. Year 2 research answered two research questions:

- How do students' claims, justifications, certainty qualifiers, and certainty rationale contribute to the overall measurement of the scientific argumentation construct?
- How do students' scientific argumentation performances change before and after each HAS investigation?

Research Design

Item Design and Early Test Design

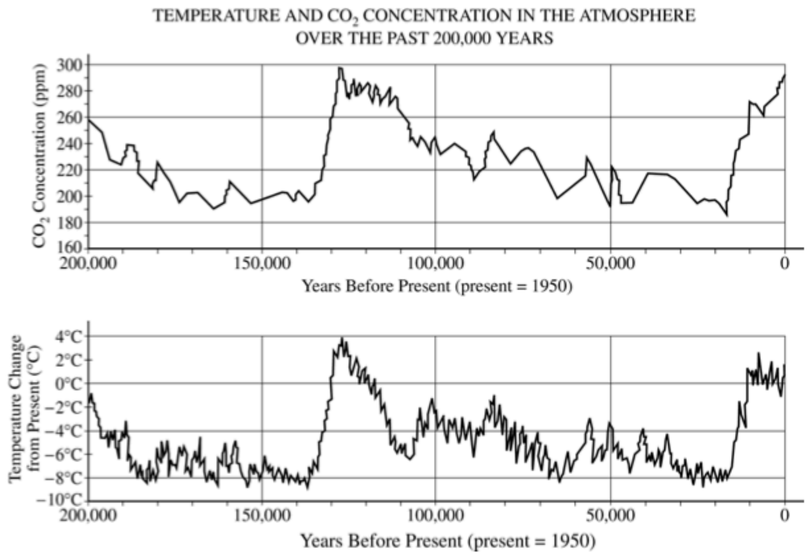
We developed item sets for two investigations: "Modeling Climate Change" and "Is there life in space?" The following topics were used as item contexts for the "Modeling Climate Change" item sets.

- Pinatubo item set: describing how Mountain Pinatubo eruptions impacted global temperatures;
- T2050 item set: predicting the temperature of 2050 based on the ice core records of global temperatures and atmospheric CO_2 levels between 125,000 years prior to 1950 and 2000;
- Ocean item set: predicting the trend of atmospheric CO_2 level when ocean temperature increases.

For the "Is there life in Space?" topic, these contexts were chosen:

- Galaxy item set: predicting a possibility of finding extraterrestrial life based on the number of galaxies and stars observed in the Universe;
- Life item set: predicting existence of Earth-like life forms by comparing information between an imaginary planet called Athena and the Earth;
- Spectra item set: predicting conditions between Uranus and Neptune based on absorption spectra.

For each of these six current science contexts, we put together four items consisting of making scientific claims (claim), explaining scientific claims based on evidence (justification), expressing the level of certainty about explanations for the claims (uncertainty), and describing their source of uncertainty (conditions of rebuttal). For claims, either multiple-choice or short-answer item format was used. For justifications, we provided data in graphs, tables, or written statements and asked students to "Explain your answer" in an open-ended format. Then, students were asked to rate their certainty on a five point Likert scale from "1" being not certain at all to "5" being very certain. Students were then asked to explain their ratings. A scientific argumentation item set called T2050 item set is shown below.



The graphs show the variation of carbon dioxide concentration and air temperature in Antarctic ice cores over 200,000 years (left side) before 1950 (right side). The upper graph shows carbon dioxide concentration in parts per million (ppm). The lower graph shows the change in air temperature.

(Source: 2006 Environmental Science AP Exam)

The CO2 concentration in the year 2000 was measured at 370 ppm. Scientific models predict that atmospheric CO2 will increase to 500 ppm in the year 2050. Based on the trends in the graphs, how much will the air temperature change between 2000 and 2050?

CLAIM
Will the temperature be higher or lower in 2050?
- higher
- lower
- no change

How many degrees will the temperature change? ______

EXPLANATION
Explain how you made your prediction.

UNCERTAINTY
How certain are you about your prediction for the air temperature in 2050?
(1) not certain at all
(2)
(3)
(4)
(5) very certain

UNCERTAINTY RATIONALE
Explain what influenced your uncertainty in question #7.

Data Collection and Analysis

In the in the early part of the 2010-2011 school year, the test containing the six scientific argumentation item sets was administered online to a total of 956 Earth science students taught by 12 teachers in six middle and high school schools located in the Northeastern United States. Among the students, 52% were female; 90% spoke English as their first language; 83% were middle school students; and 70% used computers regularly for homework. It took about 30 to 40 minutes for students to complete the test. We eliminated students who did not complete more than 50% of the 24 items to ensure the accuracy of the ability estimates. As a result, 837 students were included in the analysis.

Data Coding

- Claim items were dichotomously coded, "1" for claims that were consistent with what current scientists would claim and "0" for claims that were not.
- Explanation items were coded based on whether scientifically relevant evidence or relevant pieces of knowledge was included and how well students coordinated between knowledge and evidence. See Table 2 for an example of a scoring rubric on the justification item in the Spectra item set. See Figure RA3.
- Certainty items were coded as follows: "1" and "2" responses were assigned to uncertain (score 0), "3" to neutral (score 1), and "4" and "5" to certain (score 2) categories.
- Student responses to conditions of rebuttal items were assigned to four levels: No information (score 0), personal (score 1), scientific within investigation (score 2), scientific beyond investigation (score 3). See Table RA2.

Data Analysis

We used descriptive statistics to show what types of scientific claims, justifications, uncertainty levels, and conditions of rebuttal students exhibited. Since claim items were scored from 0 to 1, justification items from 0 to 4, uncertainty items from 0 to 2, and conditions of rebuttal items from 0 to 3, we used the Rasch Partial Credit Model (Rasch, 1966) shown below to fit the data (PCM; Wright & Masters, 1982):

$$P_{nix}(\theta) = \frac{\exp[\sum_{j=0}^{x}(\theta_n - \delta_i - \tau_{ij})]}{\sum_{r=0}^{m_i}[\exp\sum_{j=0}^{r}(\theta_n - \delta_i - \tau_{ij})]}$$

where $P_{nix}(\theta)$ stands for the probability of student *n* scoring *x* on item i . θ stands for the student location on the knowledge integration construct in this study. δ_i refers to the item difficulty. τ_{ij} (*j* = 0, 1, ..*m*) is an additional step parameter associated with each score (*j*) for item *i*.

The scientific argumentation construct in the HAS project addresses content understanding through claim and justification as well as uncertainty of the claim given evidence and reasons for uncertainty as conditions of rebuttal. Therefore, the scientific argumentation construct is a broader and more extensive construct than understanding of content and does portray more authentically scientific argumentation as performed by scientists. We created two tests for the HAS curriculum study: one for the climate investigation and another for the space investigation. For each investigation, the posttest was longer than the pretest and two identical scientific argumentation item sets appeared in both tests. The full item content for two tests are shown in

Table 6. Individual students took an online pretest before the investigation was implemented and an online posttest right after completing the investigation. Student responses to pre-posttests were scored in the same way as the Assessment Validation study described above. Student performances on the identical items were compared to estimate student gains before and after the investigation.

Table 6. Item Content for early year test and pre-post tests with Climate and Space investigations

	Early Year	Climate: Pretest	Climate: Posttest	Space: Pretest	Space: Posttest
Pinatubo.C	X	X	X		
Pinatubo. J	X	X	X		
Pinatubo. U	X	X	X		
Pinatubo. R	X	X	X		
T2050.C	X	X	X		
T2050.J	X	X	X		
T2050.U	X	X	X		
T2050.R	X	X	X		
Ocean.C	X		X		
Ocean.J	X		X		
Ocean.U	X		X		
Ocean.R	X		X		
Galaxy.C	X			X	X
Galaxy.J	X			X	X
Galaxy.U	X			X	X
Galaxy.R	X			X	X
Life.C	X			X	X
Life.J	X			X	X
Life.U	X			X	X
Life.R	X			X	X
Spectra.C	X				
Spectra.J	X				
Spectra.U	X				
Spectra.R	X				
Areas.C		X	X		
Carbon Cycle.MC		X	X		
Average Global Temperature.MC		X	X		
CO2-Infrared Graph.MC			X		
Positive Feedback.MC			X		
CO2-Water Vapor.EXP			X		
1year-5year.MC			X		
Galaxyredshift. MC				X	x
Ogle.MC				X	X
Velocityplanet.MC				X	X
Lightintensityplanet.MC					X
Velolightplanet.C					X
Velolightplanet.J					X
Velolightplanet.U					X
Velolightplanet.R					X
spetraemission.EXP					X
Elliptical.MC					
Overall					

• Claims	6	5	9	5	7
• Explanations/Justifications	6	2	4	2	4
• Uncertainty/Conditions of rebuttal	6	2	3	2	3

Year Three: HAS Curriculum Study and Learning Trajectory Study

Research Questions

- How do students' scientific argumentation performances change before and after three HAS investigations? How consistent are student performance changes across teachers? How are students' performance changes correlated with their gender, technological experience, and ELL (English Language Learner) status?
- To what extent do students who learned with HAS investigations make progress on scientific argumentation between the beginning and the end of a school year across teachers?
- How does students' scientific argumentation progress throughout the year across teachers?

Research Design

Instrument Design. The Year 2 Assessment Validation Study confirmed that the scientific argumentation assessment approach was working conceptually and psychometrically. However, the number of scientific argumentation items used for each investigation was too limited and sometimes did not align well with the curriculum content. Therefore, we increased the number of scientific argumentation items sets from two to three for each investigation and more closely aligned the item content with the curriculum content. As a result, we used nine scientific argumentation item sets in the early year and end-of-year tests: three addressing climate investigation content, three addressing water investigation content, and three addressing space investigation content. For pre-post tests for each curriculum investigation, the three scientific argumentation items that appeared in the early year and the end-of-year tests were included. In addition, additional multiple-choice claim items were included in investigation-specific pre-post tests. See Table 7 for item content.

Table 7. Test Item Content and Test Administration Information

	Early year test	Climate pre-post tests	Water pre-post tests	Space pre-post tests	End year test
Albedo argumentation item set	X	X			X
T2050 argumentation item set	X	X			X
Ocean argumentation item set	X	X			X
Galaxy argumentation item set	X		X		X
Life argumentation item set	X		X		X
Planet argumentation item set	X		X		X
City water argumentation item set	X			X	X
Well argumentation item set	X			X	X

Sediment argumentation item set	X			X	X
Additional items	--	7 claim & 1 explanation items	6 claim items	5 claim & 1 explanation items	--
Total score	118	51	45	48	118
No. of students who took the test(s)	993	406	380	245	473
No. of teachers who administered the test	12	9	9	7	9

Data Collection and Analysis

Table RA4 shows when 12 teachers implemented the early year and the end of the year tests as well as three HAS investigations. All 12 teachers administered the early year test between September and October. Teacher 2 had the second cohort and administered the early year test in January for that cohort. Nine teachers administered the end of the year test. Seven of them did it in May and June. Two teachers, T3 and T6, administered the end year test in January which was after their second HAS investigation was implemented. Three teachers implemented one HAS investigation during the school year while six teachers implemented two HAS investigations. Three teachers implemented three HAS investigations. The three HAS investigations were implemented at different times during the school year because teachers chose implementation times according to their teaching schedules. The investigation implementation sequence is shown in the second column of Table 8.

Table 8. Investigation and Assessment Implementation Schedule in Year 3

Teacher Code	Investigation Sequence	No. of investigations	Early year	Water	Climate	Space	End year
T1	WCS	3	Oct	Oct	Jan	Apr	May
T2a	WS	2	Sep	Sep		Apr	May
T2b	WC	2	Jan	Jan	May		May
T3	CS	2	Oct		Oct	Dec	Jan
T4	SW	2	Oct	Jun		Jan	June
T5	WC	2	Mar	Mar	May		
T6	WC	2	Sep	Nov	Jan		Jan
T7	C	1	Sep		Oct		
T8	SW	2	Sep	Mar		Oct	May
T9	WCS	3	Oct	Jan	Mar	Apr	May
T10	SCW	3	Sep	Apr	Feb	Oct	May
T11a	WC	2	Sep	Sep	Dec		
T11b	C	1	Sep		Apr		May
T12	S	1	Sep			Nov	

Note. C=Climate Investigation, W=Water Investigation, S=Space Investigation

Items in the tests were scored similarly to the items in the Year 2 tests were scored. Claim items were scored as “1” for correct and “0” for incorrect. Explanation items were scored from 0 to 4 according to the explanation rubric illustrated in Figure 3. Uncertainty rating items were coded from 1 (very uncertain) to 5 (very certain). Uncertainty rationale items were scored from 0 to 3 as illustrated in Table 5. Maximum possible scores were 118 for the annual pre-posttests, 51 for climate investigation pre-post tests, 45 for water investigation pre-post tests, and 48 for space investigation pre-post tests.

To compare whether and how much students changed in their scientific argumentation before and after a HAS investigation, we created a total test score as well as sub scores for claim, explanation, uncertainty rating, and uncertainty rationale items. We then applied repeated measures ANCOVA. The dependent variable was the total scientific argumentation test score and the independent variable was the teacher. Students’ gender (male vs. female), technology experience (used technology for learning vs. not used), and ELL status (English as first language vs. second) were entered as covariates. We also obtained an investigation completion ratio for each teacher as an indicator for fidelity of implementation. We computed a correlation between the investigation completion ratio variable and the effect size variable. We also compared investigation completion ratios among teachers across investigations.

To compare students’ yearly progress on scientific argumentation across teachers, we applied repeated measures ANCOVA where the dependent variable was the total test score on the early year and the end year scientific argumentation tests and the independent variable was the teacher. Students’ gender, technology experience, and ELL status were entered as covariates.

We used repeated measures ANOVA to examine students’ scientific argumentation trajectories for each of the three science topics addressed in the three HAS investigations. For the climate trajectories, we used students’ scores on the three climate scientific argumentation item sets that appeared in the early-year, before and after the climate investigation, and the end-year tests. The three scientific argumentation item sets for each topic were taken by students four times over the year. For the water trajectories, we used students’ scores on the three water scientific argumentation item sets. For the space trajectories, we used students’ scores on the three space scientific argumentation item sets. The maximum possible scores were 39 for the three water item sets, 39 for the three space item sets, and 40 for the three climate item sets. To examine whether there was a systemic difference across teachers, we used the teacher as an independent variable in the repeated measures ANOVA.

2. DESCRIBE THE MAJOR FINDINGS RESULTING FROM THESE ACTIVITIES.

Year One: Characterizing Uncertainty and Scientific Argumentation Design Study

Finding 1: Duality of students' uncertainty is observed in formulating their scientific arguments: one concerning their personal knowledge, ability, and experience and the other concerning limitations of current science or investigation. Students must overcome their perceived lack of knowledge, ability, and experience in order to consider scientific limitations in formulating scientific arguments.

Though scientific argumentation has been used in science curricula and assessed in classroom settings, uncertainty involved in students' scientific arguments has not been validated psychometrically. We characterized how students explained their uncertainty rating by applying a phenomenographical approach. In that approach, coding categories were generated to accommodate all students' open-ended explanations and, as a result, coding criteria and coding hierarchies emerged inductively. We identified 13 distinct categories of student responses, shown in Table RF1. We further reduced these phenomenological codes into four numerical codes that represented (score "0") no information, (score "1") personal, (score "2") scientific uncertainty within investigation, and (score "3") scientific uncertainty beyond investigation.

Table RF1. Certainty Rationale Coding

	Source of Uncertainty	Description of Categories
No Information (Score 0)	• No response	• Did not respond to the related uncertainty item but answered the linked claim and explanation items
	• Simple off-task responses	• Wrote "I do not know" or similar answers • Provided off-task answers
	• Restatement	• Restated the scientific claim made in the claim item • Restated the uncertainty rating.
Personal (Score 1)	• Question	• Did/did not understand the question
	• General knowledge/ability	• Did/did not possess general knowledge or ability necessary in solving the question • Did/did not learn the topic (without mentioning the specific topic) • Can/cannot explain/estimate • Used data/graph/trend (without mentioning specific data patterns or factors or interpretations used in the study)
	• Lack of specific knowledge/ability	• Did not know specific scientific knowledge needed in the item set • Mentioned specific science topics or knowledge based on misconceptions
	• Difficulty with data	• Did not make sense of data provided in the item
	• Authority	• Mentioned teacher, textbook, and other authoritative sources

Scientific-Within Investigation (Score 2)	• Specific knowledge	• Referred to/elaborated a particular piece of scientific knowledge directly related to the item
	• Data	• Referred to a particular piece of scientific data provided in the item
Scientific-Beyond Investigation (Score 3)	• Data/investigation	• Recognized the limitation of data provided in the item and suggested a need for additional data • Mentioned that not all factors are considered
	• Phenomenon	• Elaborated why the scientific phenomenon addressed in the item is uncertain
	• Current science	• Mentioned that current scientific knowledge or data collection tools are limited to address the scientific phenomenon in the item

Examples of uncertainty rationales within "Personal":

- I am not sure if I made the right observations. I might of got confused on the graphs which might have caused my answer. This why I am not sure of my prediction and answer.
- I didn't understand the question.

Examples of uncertainty rationales within "Scientific uncertainty within investigation featured in the item":

- This graph is sort of confusing to make an estimate out of because there are two of them and they explain factors that I don't completely understand. Also, the patterns haven't been happening enough times to make an accurate prediction.

Examples of certainty rationales within "Scientific uncertainty beyond investigation featured in the item":

- I am not so certain because the temperature may drop if people stop letting carbon dioxide into the atmosphere.
- While I am relatively confident about my answer, I also feel like the graph spanned over such a large range -- 200,000 years -- that determining the change within one hundred years is rather difficult.

According to this four level numeric coding scheme*, 43% of the students did not provide information about the source of their certainty, 41% provided personal reasons for their certainty, and 16% provided scientific explanations,* both within and beyond the investigations, for their certainty. These results indicate that status quo students in general were not used to addressing certainty in formulating scientific arguments. Most of those who did provide their reasons addressed *lack of confidence in their personal knowledge, experience, and ability, different from conditions of rebuttal for scientific uncertainty, such as lack of knowledge, theory, or equipment at the scientific community level.* What is encouraging was that about 16% of the students could address conditions of rebuttal related to frontier science without particular instruction on the science content or on the scientific argumentation.

Finding 2. Claim, explanation, uncertainty rating, and uncertainty rationale can contribute to the measurement of the scientific argumentation construct. Relationships among these four elements were explored.

First, the distributions of students' explanation levels across five item contexts were significantly different, $\chi^2(16) = 42.14$, $p < .001$. Across items, 60% or more students wrote non-normative ideas or irrelevant responses. The differences mainly occurred to the distributions of students writing irrelevant responses and no-link scores while the distributions of students receiving partial, full, and complex-links were relatively consistent across items. Second, the distributions of students' uncertainty ratings were significantly different across items, $\chi^2(8) = 30.0$, $p < .001$. The temperature prediction item, based on the temperature trend over the last 160,000 years, was rated as most uncertain by the students. The percentage of students who were uncertain ("1" and "2") was significantly higher than that of students who were certain ("4" and "5") across items. Third, students were more likely to be uncertain about their claims and justifications when they cited personal reasons for uncertainty rationale while students were more likely to be certain when they cited scientific reasons. χ^2 (4) = 62.7, $p <. 001$.

According to Rasch PCM analyses, the person separation reliability was 0.74 for Test A and 0.63 for Test B. Cronbach alpha values were slightly higher, 0.77 for Test A and 0.70 for Test B. We found that two multiple-choice items in Test A and one multiple-choice item in Test B were misfit items. Most of the multiple-choice items, explanation items, and uncertainty rationale items can produce a single scale conceptualized as scientific argumentation in this study (up to Level 4 in Table 1). The Wright Map in Figure RF1 shows the distributions of student ability and item thresholds (50% cumulative probability) on the same logit scale from -3 (easy item, less able student) to + 3 (difficult item, more able student). The higher the student on the scale, the more capable the student on the scientific argumentation construct. The higher the item on the scale, the more difficult the item on the scientific argumentation construct. Figure RF1 shows that multiple-choice items targeted the middle range of the scale while explanation items covered a wider range of the construct. The item threshold values for the four knowledge integration levels from 1 to 4 increased monotonically, indicating that a higher ability was needed for students to produce scientifically valid and elaborated coordination between a claim and evidence than to merely match a claim with evidence or to choose a correct scientific claim. Providing scientifically based rationale for uncertainty rating was associated with a higher ability on the construct than providing personal reasons. In addition, Figure RF1 shows that students needed at least to have scientifically normative ideas about evidence to evaluate uncertainty of their claims using scientific reasons.

Figure RF1. Wright Map: Student distributions (left) and item threshold distributions (right) are plotted on the same logit scale for Test A. Similar distributions were found with Test B.

```
           PERSON - MAP - ITEM - 50% Cumulative probabilities (Rasch-Thurstone thresholds)
                    <more>|
                          |                                                                                                 O30.4
                          |
 2                        +
                          |                                                                                                 O20.4
                          |
                          |
                          |                                                                                                 O10.4
                          |                                                                                        O30.3
                       X  |
                       X  |
 1                        +T                                                O13.2                                  O10.3
                          |  S02.1
                      XX  |  C06.1                                                                                 O20.3
                      XX T|  S01.1     S07.1
                          |S S05.1     S06.1     S09.1
                       X  |  C02.1     W03.1                                O33.2
                          |  C03.1                                          O23.2
                     XXX  |  W02.1     W06.1
 0                   XXX  +M S08.1     W01.1
                      XX  |  S04.1     W04.1     W05.1
                    XXXX S|                                       O23.1
                XXXXXXXX  |                                       O33.1
                      XX  |S C04.1     W07.1
                 XXXXXXX  |  W08.1                                                                      O10.2 O20.2
           XXXXXXXXXXXXX  |  C05.1     C07.1                                                            O30.2
                  XXXXXX  |  C01.1     S03.1
-1              XXXXXXXX M+T W09.1                                                           O30.1
                    XXXX  |
              XXXXXXXXXX  |                                       O13.1
                 XXXXXXX  |
                  XXXXXX  |                                                                  O20.1
                   XXXXX  |                                                                  O10.1
                     XXX  |
                  XXXXXX S|
-2               XXXXXXX  +
                       X  | Multiple choice     Uncertainty Rationale     Open-ended Explanation
                      XX  |    Items            Items (scored 0 to 2)      Items (scored 0 to 4)
                       X  |
                       X  |                     Score1: personal reasons    Score1: no-link
                      XX T|                     Score2: scientific reasons  Score2: partial-link
                          |                                                 Score3: full-link
                       X  |                                                 Score4: complex-link
-3                  <less>|
```

Year Two: Scientific Argumentation Validation Study and HAS Curriculum Study

Finding 1: We validated an assessment method to measure students' scientific argumentation ability consisting of claim, justification, uncertainty, and uncertainty rationale.

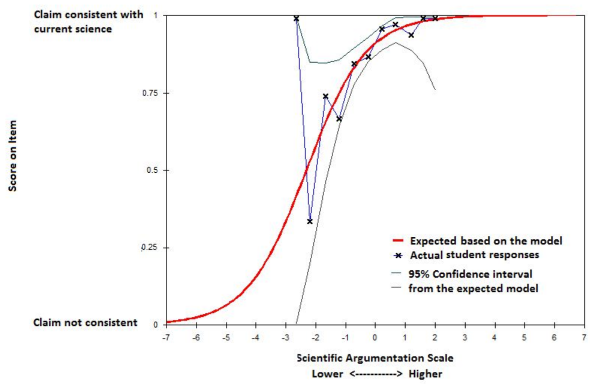
Item fit

Table RF2 shows item fit statistics in mean square values. The acceptable range for item fit to the Rasch Partial Credit Model is between 0.70 and 1.30 (Bond & Fox, 2007). There were no misfit items based on infit and outfit statistics. *According to these results, students' responses to all four types of items could be interpreted on the overall scientific argumentation scale.* Figure RF2 shows how well students' actual responses to the Life item set fit the Rasch Partial Credit Model. In all figures, the x-axis indicates students' scientific argumentation abilities from low (-7.0) to high (7.0). The y-axis represents students' scores on the item. The Rasch Partial Credit Model represents a monotonically increasing relationship between student ability and student score on the item. That is, students are more likely to receive higher scores on the item as their underlying scientific argumentation abilities increase. Students' responses to justifications, uncertainty, and conditions of rebuttal items in the Life item set closely map onto the model lines. In the claim item, this monotonically increasing relationship holds except for the very low ability students who picked the scientifically correct claim.

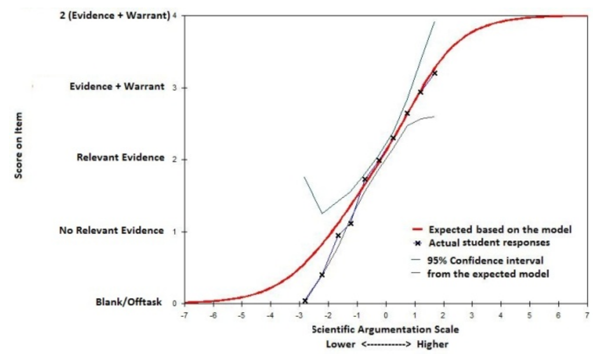
Table RF2. Rasch Partial Credit Model Analysis Results

Item			Infit		Outfit	
Items		difficulty	mean square	Error	mean square	error
(a)	Claims					
•	Pinatubo	-0.57	1.03	0.07	1.03	0.07
•	T2050	0.87	0.97	0.08	0.95	0.08
•	Ocean	1.16	1.03	0.09	1.10	0.09
•	Galaxy	-1.15	1.07	0.08	1.09	0.08
•	Life	-2.24	0.98	0.11	0.93	0.11
•	Spectra	0.20	1.00	0.08	1.00	0.08
mean item difficulty =		- 0.29				
(b)	Explanations					
•	Pinatubo	0.23	0.95	0.06	0.94	0.06
•	T2050	0.65	0.93	0.05	0.91	0.05
•	Ocean	0.10	0.94	0.04	0.94	0.04
•	Galaxy	0.01	0.97	0.05	0.97	0.05
•	Life	-0.30	0.95	0.04	0.95	0.04
•	Spectra	0.73	0.94	0.04	0.93	0.04
mean item difficulty =		0.24				
(c) Uncertainty qualifiers						
•	Pinatubo	-1.42	0.96	0.06	0.96	0.06
•	T2050	0.24	1.08	0.05	1.13	0.05
•	Ocean	-1.00	0.99	0.05	0.99	0.05
•	Galaxy	-1.38	1.08	0.06	1.18	0.06
•	Life	-1.29	0.97	0.06	0.97	0.06
•	Spectra	-0.07	1.13	0.05	1.16	0.05
mean item difficulty =		- 0.82				
(d) Uncertainty rationale						
•	Pinatubo	0.89	1.04	0.05	1.05	0.05
•	T2050	0.88	0.95	0.06	0.95	0.06
•	Ocean	1.10	0.98	0.05	0.97	0.05
•	Galaxy	0.57	1.05	0.04	1.04	0.04
•	Life	0.72	1.04	0.04	1.06	0.04
•	Spectra	1.07	0.97	0.06	0.98	0.06
mean item difficulty =		0.87				

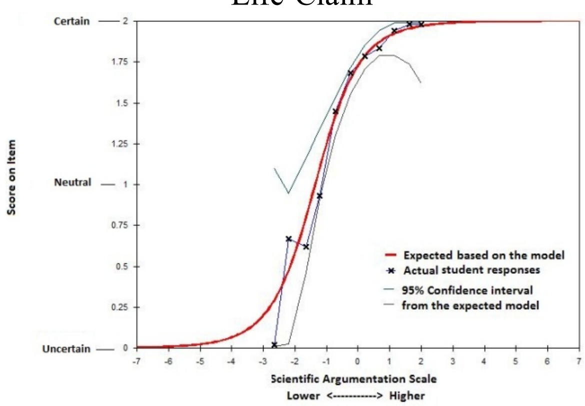
.
Figure RF2. Item characteristic curves for the Life item set.



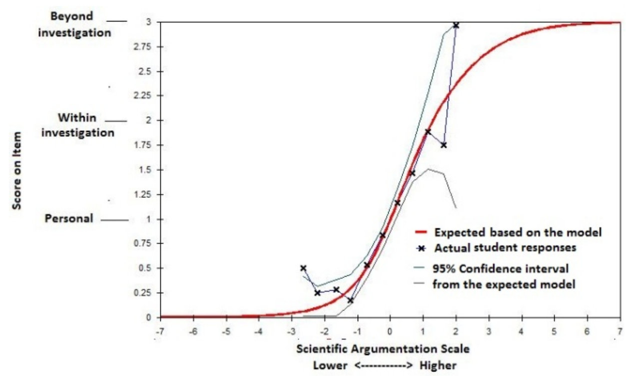
Life-Claim



Life-Explanation



Life-Uncertainty



Life-Uncertainty rationale

Rasch Scale for the Scientific Argumentation Construct

We examined how difficult each item was on the scientific argumentation scale. Table RF2 shows that the easiest item on the scale was the claim item in the Life item set with the item difficulty value of -2.24. This means that students whose scientific argumentation ability was at -2.24 had a 50% chance of answering this item correctly. The most difficult item was the claim item in the Ocean item set with the item difficulty value of 1.16. We then compared average item difficulty values across claim, justification, uncertainty, and uncertainty rationale items. The easiest item group was the uncertainty item group, followed by the claim item group. The most difficult item group was the uncertainty rationale item group. Justification items were placed between claim and uncertainty rationale items. See Table RF2. *These results indicate that the order of the required ability on the scientific argumentation scale was uncertainty → claim → justification → conditions of rebuttal, instead of the hypothesized order of claim → justification → uncertainty → conditions of rebuttal.*

Figure RF3 shows how items and students distributed on the scientific argumentation scale expressed in logit values ranging from -4.0 to +4.0. On the left side, the distribution of students according to their scientific argumentation ability is shown. *The higher on the scale, the more able students are on the scientific argumentation construct.* On the right side, item thresholds of all scores in claim, justification, uncertainty, and uncertainty rationale items are shown. An item threshold is defined as students with the matching ability would have a 50% chance of receiving a score *j* as compared to receiving a score *j – 1*.

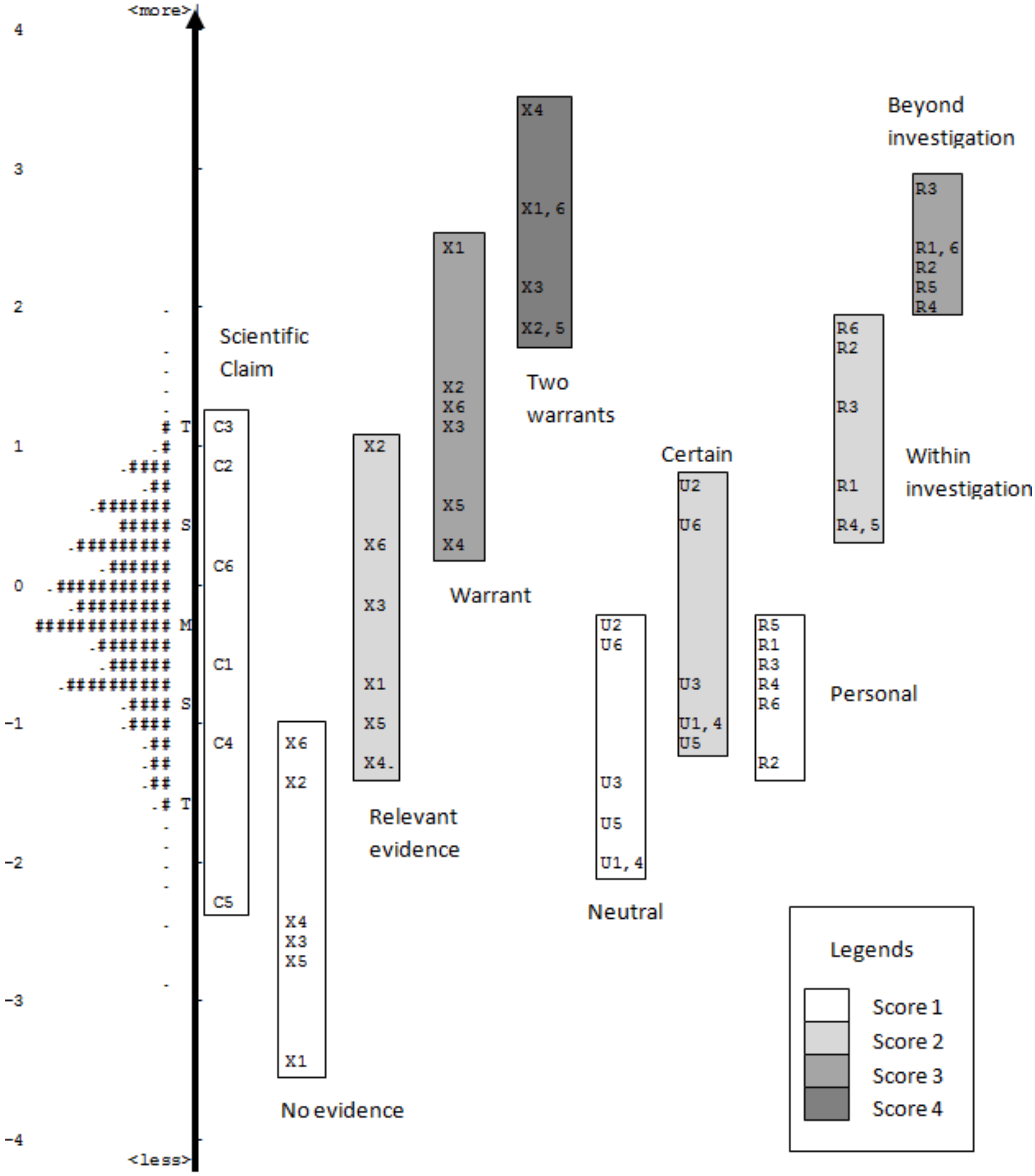
The locations of these item threshold values across four types of items were grouped in bars. The explanation items covered the widest range of the scientific argumentation scale between -3.60 to +3.80. Uncertainty rationale items covered the range of -1.35 to +3.10. The range covered by claim items was smaller than the range covered by explanation and uncertainty rationale items but slightly larger than the range covered by the uncertainty items. Both uncertainty and claim items covered the middle range of the scientific argumentation scale.

The item threshold band of making single warrants in explanations was located at a similar range to that of explaining uncertainty within investigation. The band of making two or more warrants was located at a similar range to that of explaining uncertainty beyond investigation. *These findings suggest that students who could make single warrants were more likely to consider conditions of rebuttal within investigation. Students who could make multiple warrants were more likely to consider conditions of rebuttal beyond investigation, indicating that students need to make multiple warrants based on multiple evidence pieces in order to consider limitations of the investigations imposed by current science, inquiry method, or other factors.*

The scientific argumentation scale, shown in Figure RF3, had a person separation reliability of 0.77 and an item separation reliability of 1.00.

In summary, Rasch analysis results indicate that (1) students' responses to all four argumentation elements can be interpreted on a single scale, (2) higher scientific argumentation abilities are needed in the order of uncertainty rating, claim, explanation, and uncertainty rationale on the scientific argumentation scale, (3) explanation and uncertainty rationale items measure a wider range of the scientific argumentation scale than claim and uncertainty rating items, (4) students who make a single warrant are more likely to think about conditions of rebuttal within the context of investigation, and (5) students who make two or more warrants are more likely to consider conditions of rebuttal beyond the context of investigation. These results indicate that students' scientific argumentation ability can be measured with these four item elements in a psychometrically valid manner and that students' performances on tests comprising the scientific argumentation item sets can be compared on a wide variety of statistical procedures because the scale can be considered interval.

Figure RF3. Wright Map



Note. “C” = Claim; “J” = Justification; “U” = Uncertainty; and “R” = Uncertainty rationale; “1” Pinatubo Item Set; “2” T2050 Item Set; “3” Ocean Item Set; “4” Galaxy Item Set; “5” Life Item Set; “6” Spectra Item Set; “#” represents 7 students.

Finding 2: Students significantly improved their scientific argumentation ability after HAS investigations.

The first classroom version of the climate change investigation was implemented by five teachers in Year 2, and the first version of the space investigation was implemented by six teachers in Year 2. Repeated measures ANOVAs on the student argumentation variable showed a significant teacher effect, $F(4,192)=13.13$, $p<.001$ for the Climate investigation and $F(5, 165) = 10.81$, $p<.001$ for the Space investigation. This means that students' scientific argumentation abilities as a group were different across teachers. After controlling for these significant variations, we found a statistically significant overall improvement in students' scientific argumentation ability before and immediately after both investigations, $F(1, 192) = 9.71$, $p<.01$ for the Climate investigation; $F(1, 165) = 4.28$, $p<.05$ for the Space investigation. See Table RF3. Further, we found significant interaction effects between student improvement and teacher ($F(1, 192) = 4.00$, P,.01 for the Climate investigation and $F(1, 165)=4.33$, $p<.001$ for the Space investigation, indicating the amount of average student improvement differed by teacher.

Table RF3. Average scientific argumentation scores before and after HAS investigation

(a) Climate Investigation

Teachers	Pre	Post	d
T1 (n=19)	11.42	11.05	-0.09 SD
T2 (n=29)	10.28	11.10	0.32 SD
T3 (n=69)	7.20	8.36	0.40 SD
T4 (n=50)	9.82	9.64	-0.07 SD
T5 (n=30)	10.07	12.10	0.56 SD
All (n=197)	9.16	9.92	0.23 SD

(b) Space Investigation

Teachers	Pre	Post	d
T2 (n=24)	10.71	11.50	0.34 SD
T4 (n=41)	11.02	10.02	-0.41 SD
T5 (n=23)	12.48	14.13	0.67 SD
T9 (n=39)	12.85	13.38	0.21 SD
T10 (n=25)	13.28	14.32	0.33 SD
T11 (n=19)	12.79	12.32	-0.18 SD
All (n=170)	12.12	12.42	0.10 SD

Note. "d" represents Cohen's d, mean difference between pre and posttests divided by the pooled standard deviation.

Even though the improvement in students' argumentation was significant for both investigations, the amount of improvement was smaller than we anticipated. We speculated that the misalignment between scientific argumentation item contexts and what students did during the investigations played a role. For example, science contexts for scientific argumentation items used in the pre and posttests addressed static scientific data that were already collected by scientists while students were engaged in computational modeling activities during the investigations. The scientific argumentation items used in the pre and posttests addressed the outcome of climate change or the answer to the existence of extraterrestrial life, while students would engage in more complex reasoning activities that explain the scientific processes and investigations of climate change and life on other planets. Therefore, we revised scientific argumentation items and added new items to be better aligned with the activities in the investigations for Year 3.

The water investigation was developed in Year 2. The design of scientific argumentation items and investigation activities was greatly influenced by what we learned from our assessment results from year 2 implementations. Our improved designs for the investigation activities and improved alignment between assessment content and curriculum investigation resulted in a huge increase in Effect Size from pretest to posttest, for 1.7 SD for two teachers combined. See Table RF4. Repeated measures ANOVAs indicate a significant improvement in students' scientific argumentation ability, $F(1, 50) = 64.15$, $p<.001$. We also found a significant teacher effect, $F(1, 50)=12.98$, $p<.001$, indicating the two student groups were not homogeneous. A significant interaction effect between time and teacher indicate, $F(1, 50)=13.18$, $p<.001$, that teacher 1's class significantly gained more than teacher 2's class with the water investigation.

Table RF4. Student argumentation performance comparison before and after the Water investigation

Teachers	Pre	Post	d
T1 (n=36)	15.89	24.03	2.57 SD
T2 (n=16)	21.19	24.25	0.76 SD
All (n=52)	17.52	24.10	1.71 SD

Since we made necessary changes to the Climate and Space investigations, we anticipated larger student gains in scientific argumentation in the next round of Climate and Space investigation implementations during Year 3.

Year Three: HAS Curriculum Study and Learning Trajectory Study

Finding 1: Students significantly improved their scientific argumentation ability before and after all three HAS investigations. The improvement occurred in all four elements of scientific argumentation, i.e. claim, explanation, uncertainty rating, and uncertainty rationale.

Using identical pre-post tests, we assessed students' scientific argumentation ability before and after the implementation of the latest versions of the Climate, Water, and Space investigations. The pretests were taken before the respective investigations and the posttests were taken just after the investigations were finished. The pretest and the posttest of each investigation consisted of claim, explanation, uncertainty rating, and certainty rationale items. During the 2011-2012 school year, students of nine teachers completed the Climate pre-post tests, those of nine teachers completed the Water pre-post tests, and those of seven teachers completed the Space pre-post tests. Table RF5 shows descriptive statistics for student performances on four argumentation elements separately as well as combined. Student performance changes from pre to posttests are shown in Effect Size defined as Cohen's d (the mean difference from pre to posttest divided by the pooled standard deviations).

Table RF5. Student Improvement Before and After HAS Investigations

	No. of items	Maximum allowed score	Pretest Mean (SD)	Posttest Mean (SD)	Effect Size (d)***
(a) Climate Investigation (N= 448 students from nine teachers)					
Claim	10	11	5.8 (2.1)	6.6 (2.2)	0.37 SD
Explanation	4	16	5.1 (2.2)	6.1 (2.5)	0.43 SD
Uncertainty rating	3	15	9.5 (2.5)	10.9 (2.6)	0.55 SD
Uncertainty rationale	3	9	2.4 (1.3)	2.7 (1.5)	0.22 SD
Total	20	51	22.8 (5.9)	26.3 (6.5)	**0.56 SD**
(b) Water Investigation (N= 409 students from nine teachers)					
Claim	9	9	4.7 (1.8)	5.7 (1.7)	0.57 SD
Explanation	3	12	4.5 (1.9)	5.8 (1.8)	0.70 SD
Uncertainty rating	3	15	10.9 (2.6)	12.3 (2.3)	0.57 SD
Uncertainty rationale	3	9	2.7 (1.7)	3.3 (1.9)	0.33 SD
Total	18	45	22.8 (5.9)	27.1 (5.6)	**0.75 SD**
(C) Space Investigation (N=270 students from seven teachers)					
Claim	8	8	4.1 (1.5)	5.1 (1.7)	0.63 SD
Explanation	4	16	5.9 (1.8)	7.2 (2.5)	0.60 SD
Uncertainty rating	3	15	10.3 (2.2)	11.9 (2.5)	0.68 SD
Uncertainty rationale	3	9	3.2 (1.6)	3.7 (1.7)	0.30 SD
Total	18	48	23.6 (4.5)	27.9 (6.2)	**0.81 SD**

Note. SD = Standard Deviation
Effect Size = Cohen's d = Mean difference between pre and posttests divided by the pooled standard deviation of pre and posttests.
***: All pre-post changes listed in the table are statistically significant at the $p<.001$ level.

As shown in Table RF5, students significantly improved their performance on all four elements of scientific argumentation in all three investigations. When combining all elements, students' improvement became 0.64 SD for the Climate investigation, 0.77 SD for the Water investigation, and 0.85 SD for the Space investigation. Among the four scientific argumentation elements, the most improved were students' uncertainty rating and explanations while the least improved was the certainty rationale. These results indicate that the HAS curriculum investigations supported students' content acquisition as shown in the improvement in scientific claims, scientific reasoning as shown in the improvement in explanations, and consideration of limitations of given evidence as shown in the improvement in certainty rationale. These results also indicate that (1) there was a lot of room for further improvement, and (2) scaffolding should be added to the curriculum investigations to further assist students' development of scientific argumentation, in particular on how to consider and explain uncertainty associated with scientific investigations.

Finding 2. The amount of student improvement before and after the HAS investigations differed across teachers.

Students' gains in scientific argumentation before and after HAS investigations were statistically significant for all three HAS investigations. See Table RF6. The improvement was not significantly dependent upon students' gender, technology experience, and ELL status as there were no significant interaction effects of Time with Gender and Technology. This means that students improved regardless of their gender, technology experience, and ELL status for all three HAS investigations.

Table RF6. ANCOVA Results on Students' Scientific Argumentation across Three HAS Investigations

Investigation	Climate	Water	Space
(a) Within subjects effects			
Time	11.55***	9.72**	7.76**
Time x Gender	0.004	0.03	0.83
Time x English	0.59	0.35	1.14
Time x Technology	0.50	0.29	0.76
Time x Teacher	8.66***	4.69***	13.87***
(b) Between subjects effects			
Teacher	20.04***	11.08***	13.47***
Gender	2.96	2.17	0.03
English	1.87	1.30	6.62*
Technology	1.18	17.29***	0.93

There was a significant teacher effect indicating that scientific argumentation abilities were different from teacher to teacher. This was expected as students were not randomly drawn from the student population. After controlling for variations due to teacher and students' gender, ELL status, and technology experience, there was a significant interaction effect between TIME and Teacher. That is, students' improvement was significantly different across teachers. This can be better illustrated by comparing Cohen's d values (Effect Sizes) across teachers for each HAS investigation. See Table RF7. For the Climate investigation, the effect sizes varied from -0.14 SD to 1.72 SD. For the Water investigation, the effect sizes varied from 0.44 SD to 3.07 SD. For the Space investigation, the effect sizes varied from -0.09 SD to 2.15 SD. Among 25 investigation implementations, only two investigation implementations showed no significant changes: T6's Climate investigation and T3's Space investigation. Coincidentally, T3's Space investigation was implemented in December, right before the winter break, and T6's Climate investigation was implemented in January, right after the winter break. It might be possible that students were not giving their best efforts to take tests.

Table RF7. Student Gains in Scientific Argumentation across Teachers

(a) Climate Investigation

Teacher	n	Pre Mean	Post Mean	Effect Size (SD)
T1	9	22.89	31.33	1.72 SD***
T2	101	20.47	23.71	0.61 SD***
T3	21	18.52	20.86	0.49 SD**
T5	26	28.42	33.81	1.04 SD***
T6	12	24.25	23.58	-0.14 SD
T7	105	22.67	28.81	1.17 SD***
T9	56	21.04	22.59	0.24 SD*
T10	21	26.76	29.62	0.48 SD**
T11	55	26.95	28.35	0.33 SD*
Total	406	22.89	26.47	0.58 SD***

(b) Water Investigation

Teacher	N	Pre Mean	Post Mean	Effect Size (SD)
T1	8	23.25	31.63	3.07 SD***
T2	135	19.78	25.99	1.07 SD***
T4	66	25.85	28.21	0.44 SD**
T5	31	25.48	29.94	1.08 SD ***
T6	13	24.00	27.54	1.00 SD***
T8	35	23.80	26.49	0.62 SD**
T9	47	21.11	24.49	0.56 SD**
T10	24	24.83	28.33	0.75 SD***
T11	21	28.48	31.00	0.61 SD**
Total	380	22.85	27.15	0.75 SD***

(c) Space Investigation

Teacher	N	Pre Mean	Post Mean	Effect Size (SD)
T1	9	24.1	32.8	2.06 SD***
T2	27	24.0	32.5	2.15 SD***
T3	35	22.0	21.6	-0.09 SD
T4	68	24.8	30.6	1.22 SD***
T8	28	23.2	27.4	0.94 SD***
T9	52	22.0	24.1	0.42 SD**
T10	26	25.4	30.7	1.11 SD***
Total	245	23.6	27.9	0.81 SD***

We also calculated investigation completion ratios for 25 investigation implementations as shown in Table RF8. The Climate investigation included 81 curriculum prompts that required students to respond, the Water investigation included 110 prompts, and the Space investigation included 115 prompts. On average, the Climate investigation had the highest investigation completion ratio with 75%, followed by the Space investigation with 70%. The Water investigation was the least completed by students, with an investigation completion ratio of 51%. There were wide variations among the teachers who implemented HAS investigations. Student gains were moderately positively related to investigation completion ratios, $r = 0.39$, $p < .05$. That is, student gains were related to investigation completion ratios but investigation completion ratios alone could not predict student gains. Other implementation factors should be considered to better predict student gains.

Table RF8. Fidelity of Implementation Indicator: Investigation Completion Ratio

Teacher	Climate (81 prompts)	Water (110 prompts)	Space (115 prompts)
T1	95%	99%	95%
T2	64%	32%	93%
T3	83%		84%
T4		31%	62%
T5	99%	96%	
T6	74%	69%	
T7	76%		
T8		81%	83%
T9	64%	36%	51%
T10	92%	84%	79%
T11	87%	78%	
Total	75%	51%	70%

Finding 3: Students significantly improved their scientific argumentation abilities over the year.

Early in the 2011-2012 school year, around September and October, 11 teachers administered the annual scientific argumentation pretest to 993 students. Toward the end of the 2011-2012 school year, around May and June, 9 teachers administered the same annual scientific argumentation posttest to 473 students. Among these students, 406 students took both pre and posttests. For the analysis, we took the 379 students who responded to both tests. The annual scientific argumentation test was identical and consisted of nine scientific argumentation item sets: three sets addressing climate topics, three sets addressing water topics, and three sets addressing space topics. Each item set had a maximum score of 13 and thus the whole test had a maximum score of 118. Table RF9 shows mean values for the pretest and the posttest, along with student gains in standard deviation units (Effect Size, Cohen's d). Students of all nine teachers gained statistically significantly from the pretest to the posttest with an average effect size of 1.01 SD, a large impact.

We applied repeated measures ANCOVA to examine how student gains in scientific argumentation were correlated with student characteristic variables such as gender, technology experience, and ELL status and how student gains differ across teachers. According to ANCOVA results, students' argumentation performances did not significantly differ by gender, $F(1, 367) = 1.77$, $p = 0.18$, and ELL status, $F(1, 367)=1.21$, $p=0.27$. Nor were there significant performance differences by gender and ELL status in student gains in scientific argumentation. However, there was a significant technology experience effect, $(1, 367) = 5.21$, $p < .05$. Independent samples t-tests indicate that students' argumentation score was not significantly different in the pretest by their technology experience, $t(378)=1.19$, $p=0.23$, but became significantly higher for students with technology experience in the posttest, $t(378)=4.25$, $p<.001$, by an ES of 0.44 SD. In fact, the amount of student gains were significantly different between the two groups, $F(1, 367) = 9.88$, $p < .01$. These indicate that students with technology experience gained scientific argumentation abilities to a greater extent between early and end of the year than those without.

There was a significant main teacher effect, $F(8, 367) = 10.21$, $p < .001$, indicating students' scientific argumentation abilities were significantly different across teachers. See Table RF9 for differences in pretest means and posttest means across teachers. In addition, the amount of yearly gains were significantly different across teachers, $F(8, 367) = 8.52$, $p < .001$, ranging from as small as 0.23 SD to as large as 3.06 SD.

Table RF9. Descriptive Statistics for Early-Year and End-Year Scientific Argumentation Tests across Teachers

Teacher	n	Pre Mean	Post Mean	D (Effect Size) (Unit: SD)
T1	9	50.11	82.56	3.06***
T2	137	45.74	63.20	1.45***
T3	35	49.49	52.37	0.23*
T4	41	56.85	68.00	0.84***
T6	10	45.00	57.20	1.22***
T8	40	55.55	66.88	1.13***
T9	49	48.43	58.67	0.76**
T10	24	55.29	72.04	1.36***
T11	34	60.41	68.41	0.98***
Total	379	50.68	63.85	1.01***

Finding 4: Students' scientific argumentation trajectories indicated improvement over time. The largest improvement coincided with the implementation of HAS investigations. Students retained or even further improved their scientific argumentation after HAS investigations were finished.

Figure RF4 shows mean plots for students' scores on the three Climate scientific argumentation item sets across four time points. On average, students significantly improved over time, $F(3,636) = 76.33$, $p<.001$. The largest improvement coincided with the Climate investigation implementation time. Students improved their scientific argumentation between the beginning of the school year and before the Climate investigation probably because students learned climate topics or related science prior to using the investigation. Students also well retained their scientific argumentation abilities after the Climate investigation was finished. There was a significant teacher effect, $F(6,212) = 10.69$, $p<.001$, indicating students were not sampled from the homogeneous student population. The student trajectories were significantly correlated with the teacher variable, $F(18, 636)=4.46$, $p<.001$. Figure RF4 indicates this interaction effect, as the trajectories were not identical across teachers. In particular, T6 and T9 departed the pattern of small improvement→ larger improvement → maintain. The T6 trajectory showed a statistically non-significant decline before and after the investigation. The T9 trajectory shows a decline between after the investigation and at the end of the year. The T1 trajectory showed the largest improvement before and after the Climate investigation.

Figure RF4. Climate Student Trajectories by Teacher

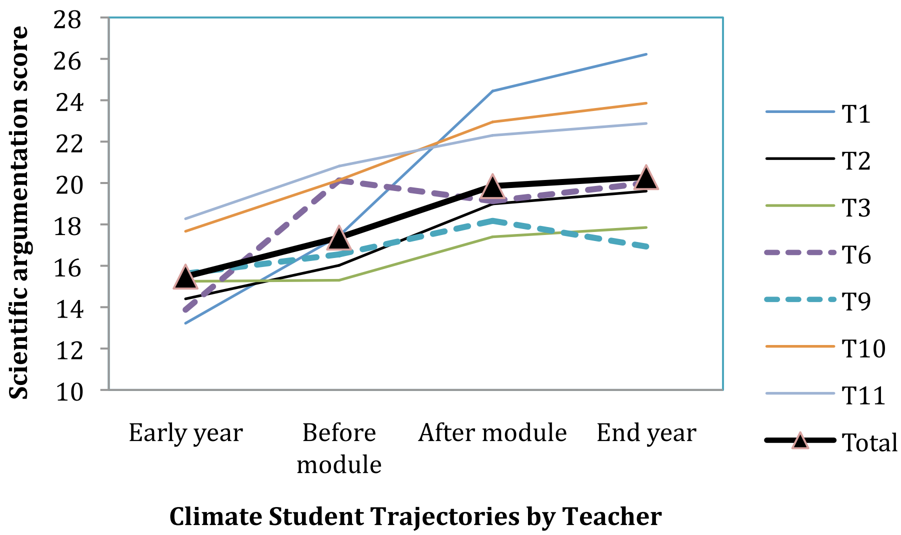


Figure RF5 shows student trajectories for the Water investigation. The general pattern was that students improved between the beginning of the year and prior to the Water investigation, followed by a grater improvement before and after the Water investigation. Then, students' scientific argumentation scores slightly declined. The student improvement over time was statistically significant, $F(3, 741) = 83.42$, $p<.001$. There was a significant teacher effect as in the Climate trajectories, $F(6, 247) = 6.76$, $p<.001$. Figure RF5 shows statistically significant variations across teachers in student trajectories, $F(18, 741)=2.35$, $p < .01$. The most distinguished trajectory was T6's as there was a noticeable drop between after the investigation was finished and the end of the school year. Again, the T1 trajectory shows the largest improvement before and after the investigation.

Figure RF5. Water Student Trajectories by Teacher

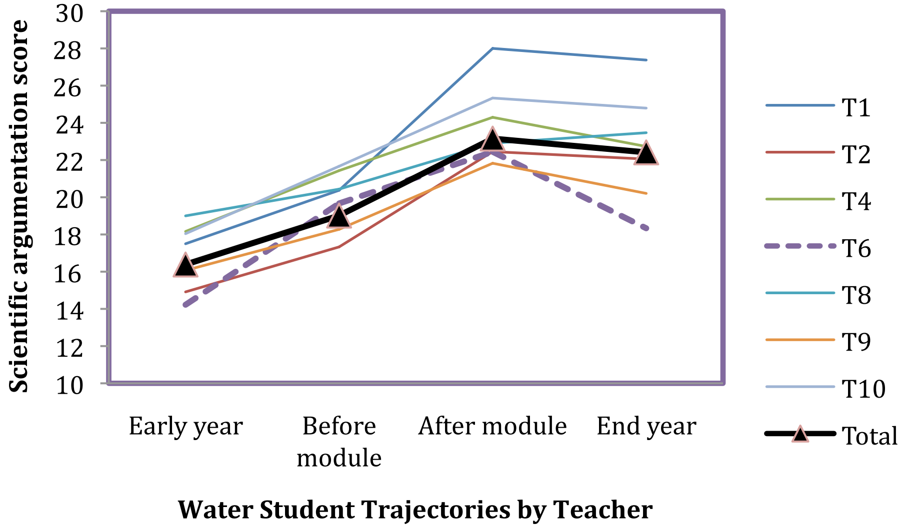
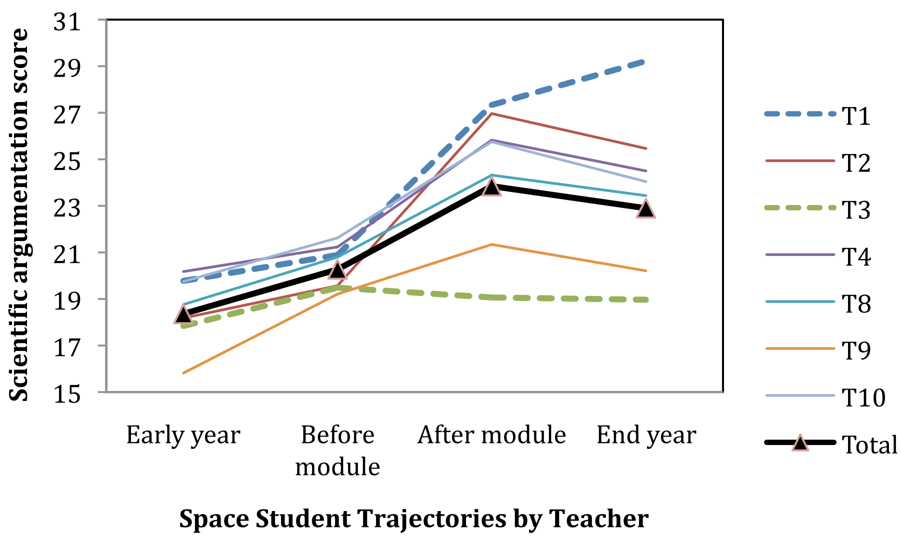


Figure RF6 shows student trajectories for the Space investigation. The general pattern of overall improvement was similar to the general pattern found with the Water investigation. The improvement was statistically significant over time, $(3, 546)= 115.72$, $p<.001$. There was a significant teacher effect, $F(6, 182) = 10.16$, $p<.001$. The student trajectories were significantly different across teachers, $F(18, 546) =5.68$, $p<.001$. In particular, T3 shows almost no changes across time points. As discussed earlier, T3 implemented the Space investigation just before the winter break when students might not take the Space investigation seriously and administered the end year test after the winter break when students might not remember what they learned before the winter break. The T1 trajectory indicates the largest improvement before and after the Space investigation, followed by a further improvement towards the end of the year.

Figure RF6. Space Student Trajectories by Teacher



Finding 5: Students' gains before and after HAS investigations were much greater with the revised versions implemented in year 3 than with the ones implemented in year 2.

Table RF10 shows that students made a greater pre-post test improvement in their scientific argumentation abilities with the Year 3 versions of the Climate and the Space investigations than the Year 2 versions.

Table RF10. Student Gains by Curriculum Version

	First version (Year 2)		Second version (Year 3)	
	Student n (Teacher n)	Mean ES	Student n (Teacher n)	Mean ES
Climate investigation	199 students (5 teachers)	0.17 SD***	406 students (9 teachers)	0.58 SD***
Space investigation	173 students (6 teachers)	0.05 SD***	245 students (7 teachers)	0.81 SD***

References

Bond, T. G., & Fox, C. M. (2007). Applying the Rasch model. Mahwah, NJ: Lawrence Erlbaum Associates.

International Association for the Evaluation of Educational Achievement (IEA) (1995). *TIMSS science items: Released set for population 1 (third and fourth grades)*. Chestnut Hill, MA: Boston College.

Joint Committee for Guides in Metrology. (2008). International vocabulary of metrology-basic and general concepts and associated terms. Geneva, Switzerland: International Organization for Standardization.

Kennedy, D. & Norman, C. (2005). 125 Questions: What don't we know? Science, sciencemag.org/sciext/125th/.

Lederman, N. G., & O'Malley, M. (1990). Students' perceptions of tentativeness in science: Development, use, and sources of change. Science Education, 74(2), 225-239.

Lindley, D. (2007). Uncertainty: Einstein, Heisenberg, Bohr, and the struggle for the sole of science. New York: Anchor Books, A Division of Random House, Inc.

Rasch, G. (1966). Probabilistic models for some intelligence and attainment tests. Chicago, IL: University of Chicago Press.

Toulmin, S. (1958). The uses of argument. New York: Cambridge University Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago, IL: Mesa Press.

Wu, M., Adams, R., & Wilson, M. (1997). ConQuest: Generalized item response modeling software. Melbourne, AU: ACER Press.