EARTH SCIENCE ASSESSMENT WITH AUTOMATED FEEDBACK (ESAAF)

The *Earth Science Assessment with Automated Feedback* (ESAAF) project is a **Design and Development Project** submitted to the Assessment Strand of the DRK-12 program. ESAAF plans to make critical improvements on current automated scoring models to accommodate formative earth science assessments with complex scoring rubrics and provide immediate feedback to support the teaching and learning of argumentation related to climate change and fresh water availability in secondary school classrooms.

IMPORTANCE & PROJECT GOALS

Importance of the science content chosen. Earth and Space Sciences (ESS) are one of the key content areas contributing to science literacy as highlighted in *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*, (NRC, 2012) and the *Next Generation Science Standards* (NGSS; NGSS Lead States, 2013). Topics such as climate change and the future of fresh water availability represent a complex set of interactions that blend science and human impacts. Climate change has drawn world-wide concerns, and recent extreme weather events such as *Hurricane Sandy* and tropical storm *Irene* are likely results of global warming (Ritter, 2012). It is now considered certain (>95%) that human influence has been the dominant cause of the warming (Intergovernmental Panel on Climate Change [IPCC], 2013). Therefore, it is of critical importance to educate students with climate science literacy to help them make informed decisions about actions that affect the climate (U.S. Global Change Research Program, 2009). Fresh water availability is another focus of IPCC (2007). According to the International Water Management Institute (IWMI), approximately 1.2 billion people live in areas with physical water scarcity (i.e., dry areas), and another 1.6 billion people face economic water shortage with water use growing at a faster rate than the rate of the population increase (Molden, 2007). The situation is further exacerbated by the unpredictable distribution of water caused by climate change (Rogers, 2008).

Earth science curricula and assessments should focus on how human activities have altered environments and their impacts on living things (e.g., ESS3.D Global climate change in the new standards). The new standards urge these topics to be taught in order for future citizens to make scientifically informed decisions about the consequences of human actions. A major challenge facing Earth science educators is how to develop and support rich science curricula with this new focus on core ideas. In ESAAF, students will learn about human-Earth interactions that include core concepts such as the water cycle to understand fresh water distribution, and the atmospheric greenhouse effect to understand climate change. Students will also develop scientific argumentation skills that enable them to make sense of scientists' data and models to explain the impact of human actions on Earth's systems (Ledley et al., 2011).

Importance of scientific argumentation. Scientific argumentation is identified as one of the eight scientific practices important for K-12 students in the NGSS (NGSS Lead States, 2013). Scientific argumentation involves making a claim from evidence about a scientific question based on the generally accepted scientific knowledge and research framework, and explicitly addressing boundaries of the claim's application (Toulmin, 1958). Engaging students in scientific argumentation can help students deepen the learning of science concept (Aufschnaiter et al., 2008), incorporate science epistemology into their learning (Jimenez-Aleixandre et al., 1999; Sandoval, 2003), engage in collaborative learning (Chinn & Osborne, 2010; Duschl & Osborne, 2002), and support decision-making (Kuhn & Udell, 2003). One aspect that has been overlooked in science education research, however, is how students treat uncertainty in formulating their scientific arguments (Bricker & Bell, 2008). As no scientific evidence can support a claim with 100% accuracy, any scientific argument has a potential for rebuttal. Therefore, in learning about the practice of scientific argumentation, students should not only need to coordinate evidence with relevant scientific theory but also need to properly address sources of uncertainty inherent in their scientific arguments. Uncertainty in students' argumentation could stem from students' confidence in their understanding of knowledge and their ability to perform investigations (Metz, 2004). Uncertainty in students' argumentation can also resemble scientists' treatment of uncertainty focusing on limitations of current scientific knowledge base, experiments, equipment, and models (Lee et al., in press). Building on prior research (see Prior Results), we will include four components in assessing argumentation, such as claims, explanations, student rating of uncertainty, and uncertainty rationale, which are described in more details in the assessment section.

Importance of automated scoring of items with complex scoring rubrics. It is agreed among researchers that constructed-response items allow for a greater degree of construct representation and are more authentic than multiple-choice items (Lane, 2004; Lee et al., 2011; Shepard, 2000). However, as scoring constructed-response items is both costly and time-consuming (Wainer & Thissen, 1993), the use of constructed-response items is limited in both large-scale and classroom assessments. For example, national and international comparison assessments such as the NAEP and TIMSS largely make use of multiple-choice items due to efficiency and scoring considerations (Quellmalz et al., 2005; Singer et al., 2006). Automated scoring could serve as a viable solution to scale up the use of constructed-response items. In addition to saving cost and improving efficiency, automated scores can also help reduce errors and biases typically introduced by human raters (Williamson, et al., 2012; Zhang, 2013).

In the past three decades, automated scoring has been applied in a wide range of scoring domains, such as writing quality (Burstein & Marcu, 2002), mathematics (Sandene et al., 2005; Koedinger, McLaughlin, & Heffernan, 2010), written content (Graesser, 2011), and speech (Higgins, Zechner, Xi, & Williamson, 2011). Examples of automated scoring systems for content domains include c-rater™ and c-rater-ML offered by ETS (Leacock & Chodrow, 2003), AutoMark (Mitchell et al., 2002), and Intelligent Tutoring Systems such as AutoTutor, ITSPOKE, and TLCTS (Graesser, 2011). Despite the growing application of automated scoring (e.g., Graesser, 2011; VanLehn et al., 2007), many previous uses focus on items that are either transformed from multiple-choice items or have simplified scoring rubrics (e.g., correct, partially correct, or wrong; Attali & Powers, 2008; Attali et al., 2008). Very few studies have evaluated automated scoring for items with complex scoring rubrics (Liu, et al., 2013). In ESAAF, we plan to enhance automated scoring of items with rubrics that differentiate among more than three levels of understanding. We will focus on automated scoring of items that measure argumentation in the context of climate change and fresh water availability. For the assessment results to be of diagnostic value to teachers, it is important to develop automated scoring that can accurately evaluate multiple levels of understanding. The automated scores will then be used to facilitate immediate feedback to students.

Importance of immediate, automated feedback to facilitate learning and teaching. Educational experts consider feedback a critical element to catalyze learning (Black et al., 2003; Clarke, 2003; Hattie, 2009; Sadler, 1998). Several meta-analysis studies showed that the effect sizes of feedback are moderate to large in improving learning (0.40 to 0.80 standard deviation [SD] units) (Azevedo & Bernard, 1995; Kluger & DeNisi, 1996; Shute, 2008). Research also showed that for classroom studies students benefit more from immediate rather than delayed feedback (Anderson et al., 2001; Kulik & Kulik, 1988; Shute, 2008). Immediate feedback will correct errors and misconceptions before they are encoded into students' learning (Brosvic & Cohen, 1988; Corbett & Anderson, 1989, 2001; Dihoff et al., 2003), particularly in a computer-based instruction environment (Azevedo & Bernard, 1995; Shute, 2008). Despite its appealing benefits, instant feedback on constructed-response items is unrealistic in current classrooms (Gibbs & Simpson, 2004). In addition, scoring constructed-response items takes away teachers' time from lesson planning and working with struggling students (National Council of Teachers of English, 2008).

We will investigate *when, how, for whom*, and *under what conditions* feedback is most effective. The goals of ESAAF feedback are to promote students' use of productive learning strategies, help students identify their misconceptions, and increase their persistence and experiences in learning science (Brown et al., 2012; Kluger & DeNisi, 1996). Some of the existing intelligent tutoring systems (e.g., AutoTutor, ASSISTments, DeepTutor) are conversational in nature and probe students' responses after they type one word to two sentences, while the ideal answer is about three to seven sentences (Graesser et al., 2004). We will take a different approach by allowing students to complete their answers and providing feedback based on the evaluation of the entirety of student responses. Prior research shows that interrupting feedback provided when a student is engaged in problem solving may impede learning (Shute, 2008). Our feedback will pay attention to timing, complexity, delivery, and accuracy as these factors determine the effectiveness of feedback (Black & Wiliam, 1998; Crooks, 1988; Shute, 2008;

Sadler, 2010). Our feedback is intended to provide students with the opportunity to review their answers, point them to the relevant instructional steps, and allow them to revise their responses. As an improvement to most current intelligent tutoring systems (e.g., AutoTutor), we will go beyond the individual student and provide class-level information to the teachers. We will design a synthesized report of the individual scores and feedback to give the teacher a snapshot of class-level performance. The immediate, class-level feedback is anticipated to help teachers make instructional decisions that are evidence-centered and data-driven. Additionally, we will offer professional development to promote teachers' understanding of automated scoring and feedback as part of formative assessment practices.

Project Goals. ESAAF responds to the national need for technology-enhanced, formative science assessments that support inquiry-based science teaching and learning. Both the *Smarter Balanced* and the *PARCC* consortia plan to use automated scoring to improve score efficiency (PARCC, 2012; SBAC, 2012). Guided by argumentation theories and drawing on automated scoring technologies with proven validity (e.g., Sukkarieh & Blackmore, 2009; Sukkarieh & Pulman, 2005), we will use *c-rater^{TM}* and *c-rater-ML*, two advanced automated scoring tools developed by ETS, to score constructed-response items. We will provide immediate feedback containing diagnostics of individuals and class-level performance. The following are our research and development goals:

1. Develop and validate automated scoring and feedback on formative assessments targeting climate change and fresh water availability for secondary school students;
2. Investigate ***when***, ***how***, ***for whom***, and ***under what conditions*** feedback can be effective in promoting learning along the dimensions of feedback type (e.g., content vs. epistemic; diagnostic only vs. diagnostic plus suggestive);
3. Develop professional development resources to help teachers use automated diagnostics to improve instruction and assessment practices;
4. Develop an interactive score reporting system that provides both customized individual feedback to students and class-level snapshots to teachers.

Our partners are **ETS**, a leader in automated scoring and innovative assessment, the **Concord Consortium (CC)**, a leader in developing and implementing technology-enhanced science curricula, and the **University of California, Santa Cruz (UCSC)**, which will incorporate learning theories of argumentation in the context of earth science topics.

RESULTS FROM PRIOR NSF SUPPORT

Our prior and ongoing research on computer-based resources in assessment and learning puts us in a unique position to achieve our goals. Our research team will combine innovative tools and expertise in scientific argumentation, automated scoring, assessment, technological development, and professional development in achieving the research goals for ESAAF. We will draw specifically on the following NSF-funded projects:

High-Adventure Science (HAS) (DRL-0929774. 9/15/09 – 8/31/12. $695,075. PI: Pallant). The HAS project developed and tested three online curriculum modules, including climate change, fresh water availability, and life on other planets for secondary school students to explore questions in ESS that scientists around the world are currently investigating. Each module was designed for five class periods and included interactive computational models, real-world data, and videos of scientists discussing their computational model-based research on the same questions. Building on literature on scientific argumentation (Bricker & Bell, 2008; Clark et al., 2007; Sampson & Clark, 2008), we assessed four essential parts of argumentation: claims, explanations based on the theoretical interpretation of evidence, uncertainty ratings, and uncertainty rationale as conditions of rebuttal. With this approach, we not only assessed students' scientific reasoning captured in the explanation component of argumentation but also evaluated students' ability to articulate the uncertainties in their arguments, both considered critical in developing argumentation skills (Osborne et al, 2004). Our analysis results indicate that students made substantial improvement on claims and explanations (e.g., effect sizes being 0.40 and 0.70SDs) from pre to posttests, but their improvement in uncertainty rationale, although significant, was limited (e.g., 0.25SD), which points to the need for immediate teacher intervention that can be made possible through automated scoring and feedback. We will be using the climate change and the fresh water availability modules developed during the HAS project to test automated scoring and feedback interfaces in the ESAAF project.

High-Adventure Science: Earth's Systems and Sustainability (HAS:ESS) (DRL-1220756. 10/1/12 – 1/31/16. $2.3M, PI: Pallant, Co-PIs: Lee and Norris). Based on promising results from the exploratory HAS project, the HAS:ESS project focuses on environmental science and is currently developing three additional modules for middle and high school students. The goal of HAS:ESS is to research the effectiveness of curriculum materials to reliably convey an understanding of Earth's systems and the increasing role of human interaction with those systems, while also introducing important science practices of scientific argumentation through modeling and crosscutting concepts of systems and systems thinking. This project is built in partnerships among Concord Consortium, UCSC, and the National Geographic Society.

Two design studies were conducted to investigate whether conceptual and epistemic prompts provided *before* students formulated arguments could improve their argumentation. Students were randomly assigned to two versions of the same module within a teacher, with or without prompts in argumentation. Results indicate an interaction between effects of prompts and components of argumentation. For example, for the water module, while there was no significant difference in explanation (p=.38), there was a significant difference in uncertainty rationale (p< .001). These results indicate that pre-supplied prompts might not be ideal to help students improve their formulation of arguments. Extending from this study, we will investigate if prompts provided in the form of automated feedback would work more effectively in helping students construct an argument.

Continuous Learning and Automated Scoring in Science (CLASS) (DRL-1119670, 09/01/2011-08/31/2016, $2.5 million, Liu as Co-PI in charge of automated scoring and measurement). CLASS explores automated scoring for constructed-response items, concept maps, science narratives, and graph items. For example, both *c-rater* and *c-rater-ML,* content-based automated scoring tools developed by ETS, were used to score constructed-response items. After several iterations of rubric refinement, the agreement between *c-rater* scoring and human scoring was as high as .70 for *c-rater* and .78 for *c-rater-ML* (Linn et al., 2012; Liu et al., 2013). The automated scoring engines are integrated with the WISE platform (https://wise.berkeley.edu/), a web-based science environment, to provide instant score and feedback to students (Liu et al., 2013).

For *c-rater* scoring, a model answer is identified for the question at hand. The model answer contains a set of key concepts that *c-rater* uses to evaluate students' responses. *c-rater* uses natural language processing techniques to identify alternative ways of student expressions. Based on the presence of the key concepts, *c-rater* assigns a score to students' responses following a specified scoring rule. The following is an example of how *c-rater* was used to score the item Spoon: "*A metal spoon, a wooden spoon, and a plastic spoon are placed in hot water. After 15 seconds which spoon will feel hottest? Explain your answer.*" This item was scored by human raters using a four-level rubric that rewards students' ability to make scientifically valid connections between concepts and ideas. Based on approximately 1,000 student responses, a *c-rater* analytic rubric was created consisting of six key concepts and their paraphrases (Fig 1), and a scoring rule (Fig 2). The quadratic-weighted kappa between *c-rater* and two human raters was as high as .70, higher than the agreement between two human raters after half a day of training (Linn et al., 2012).

CLASS also pilot tested the effect of automated feedback with two teachers. Among the 258 students included in the pilot study, 126 were in the teacher condition where they received feedback from the teacher on the next day, and 132 were in the *c-rater* condition where they received immediate, automated feedback. Results showed that the students in the *c-rater* condition were as likely to revisit and revise their responses (85%) as those in the teacher condition (87%). More importantly, students in both conditions made significant and comparable gains through revising their answers (Teacher condition, effect size = 0.41SD, p<.001; *c-rater* condition, effect size = 0.38SD, p<.001).

Fig 1. c-rater concepts for the item Spoon (only selected alternatives are presented due to space limit)

```
C1: The metal spoon will feel the hottest, but it will still be the same temperature
as all of the other spoons OR the metal spoon feels hotter than it actually is
C2: The metal gets hot fastest OR heat will come to it fastest OR metal conducts heat
fastest OR metal is the fastest conductor OR heat enters metal faster OR metal absorbs
heat faster
C3: Metal conducts the most heat OR more heat comes to the metal OR metal absorbs the
most heat OR metal conducts heat more easily OR metal absorbs heat more easily
C4: Metal heats up OR metal becomes hot OR metal gets hot OR metal gets hot easily OR
heat will come to the metal OR metal absorbs heat OR metal absorbs heat easily
C5:Metal attracts heat OR metal attracts the most heat Or metal attracts more heat
C6: Heat stays in the spoon longer OR metal keeps the heat for the longest time OR
heat stays in the spoon longer OR the metal conserves heat OR heat is more apparent in
a metal object
```

Fig 2. c-rater scoring rules for the item Spoon

```
4 points C1 and (C2 or C3 or C4)
3 points  (C1 and [C2 or C3 or
C4]) and C5
3 points (C1 or C2 or C3)
2 points (C1 or C2 or C3) and C5
2 points C4
1 point C4 and C5
1 point C5 or C6
1 point None
```

How ESAAF builds on all prior research. We will heavily draw on the three above-described NSF-funded projects in implementing ESAAF. We will use the curricula developed by the HAS and HAS:ESS projects, and use their validated scientific argumentation assessments to develop automated scoring and design feedback. We will also benefit from HAS and HAS:ESS's established partnerships with teachers as the teachers have the opportunity to continue implementing the same modules in ESAAF. We will draw on the results from CLASS on automated scoring to further enhance automated scoring modeling for items with complex rubrics, investigate the effects of different feedback types on improving student learning, and study how automated scores can be used to facilitate individual and class-level feedback.

RESEARCH AND DEVELOPMENT PLAN

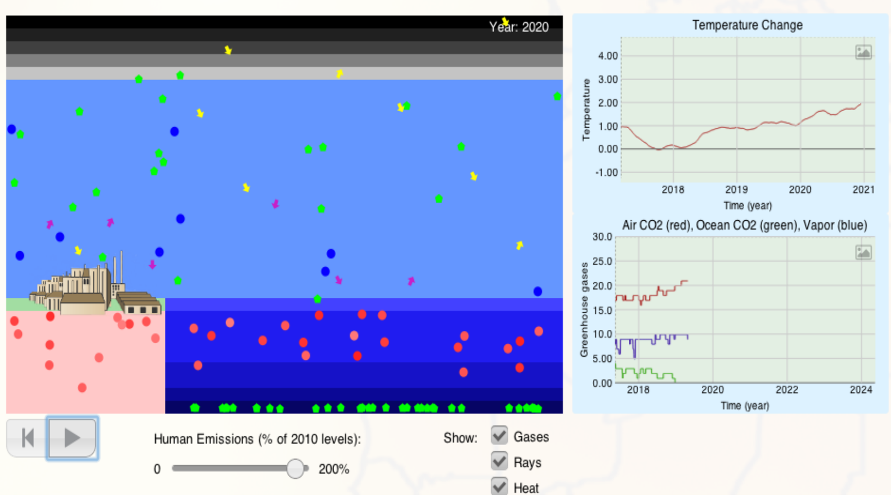
We will follow the Common Guidelines for Education Research and Development (IES & NSF, 2013) to plan our research and development. We will rigorously use three phases of research: feasibility studies, design studies, and a pilot study to investigate the effect of automated scores and feedback. Below we describe the curricula and assessments, participants, teacher professional development, research questions, and the quantitative and qualitative methods to address each question. We also discuss the development of the platform through which the automated scoring and feedback will be offered.

Curricula, Assessments, and Scoring Rubrics

As a curricular test-bed for automated scoring and feedback, we will use two online curriculum modules on climate change and fresh water availability, developed and tested in the NSF-funded HAS project (see Prior Support). These two modules were designed based on research on the use of authentic science practices in classrooms (Chinn & Malhotra, 2002; Lee & Songer, 2003), scientific argumentation (Berland & Reiser, 2009; Clark et al,., 2007; Erduran et al, 2004; Toulmin, 1958), and computational modeling (Pallant & Tinker, 2004; White & Frederiksen, 1998), and specifically addressed scientific uncertainty involved in scientists' data collection and model building (Allchin, 2012; NRC, 2012).

With the climate module, students explore factors that influence the Earth's future climate such as CO_2, albedo, volcanic activities, and human produced greenhouse gases. Students use simple climate models to explore how greenhouse gases warm the planet. Students analyze the relative effects of positive and negative feedback to make a prediction for the future of Earth's climate. In the freshwater availability module, students use models and real world data to study the water cycle and evaluate the supply and demand for freshwater in various areas of the world. They use interactive models to explore the relationships between groundwater levels, sediment permeability, rainfall, recharge of aquifers and human impact on groundwater levels. Students learn how water flows through sediments, how rates of recharge compare to rates of withdrawal, and how to assess the sustainability of water usage locally and globally.

Each module requires 5-6 class periods and includes a pretest, embedded assessments, and a posttest. Within each module, the pre and posttests are identical and include three sets of argumentation items Each scientific argumentation item set consists of a multiple-choice claim, constructed-response explanation, uncertainty rating on the five-point Likert scale, and constructed-response uncertainty rationale (see Fig 3 for a sample item). The embedded assessments in each module include eight argumentation item sets where students work with complex scientific data, simulations, and modeling to develop their arguments. All student responses to curriculum module assessments are stored electronically, which will facilitate the application of automated scoring. To assess students' scientific argumentation, in ESAAF we will continue to use the scientific argumentation item set format validated in prior studies (Lee et al., 2013; Pallant et al., 2012)**.** Prior to students' argumentation, students will either encounter data collected by scientists or computational models (Fig 3 contains a computation model).



Human Emissions (% of 2010 levels): 0 200%

Show: ☑ Gases ☑ Rays ☑ Heat

Fig 3 (left). A climate model. The yellow arrows carry a unit of energy, which is converted into heat in the earth and ocean, which then can be converted into a unit of infrared radiation, represented by purple arrows. These interact with the CO_2, represented by green dots and water vapor represented by blue dots. CO_2 is added into the environment by the slider on the bottom which changes human emissions relative to the 2010 levels of emissions.

Question: What happens if you remove all of the carbon dioxide from the atmosphere?

Claim The temperature will (a) decreases (b) increases (c) stays the same	**Uncertainty rating** How certain are you about your prediction for the air temperature in 2050? (1) not certain at all; (2), (3), (4), (5) very certain
Explanation Explain how you made your prediction.	**Uncertainty rationale** Explain what influenced your uncertainty rating.

All the constructed-response explanation and uncertainty rationale items will be scored using the following generic rubrics customized for each individual item, respectively (Tables 1 and 2), which were validated in HAS and HAS:ESS, with inter-rater reliability of .91 using both sets of rubrics.

Participants, Sample Sizes, and Power Analyses

Participants and sample sizes. We plan to recruit about 1,940 students taught by 20 teachers between the two modules over the four years of this project. The participants in the first two years will be from two schools in MA and CT. They were chosen because of their access to computers required by our research plan and proximity to the Concord Consortium for implementation support purposes. These first two schools serve a diverse student population. The Magnet School in Hartford, CT (see teacher support letter) has about 76% students of a racial/ethnic minority background and 99% of its students receive free or reduced-price lunches. The Belmont Public School in MA (see teacher support letter) serves suburban population with 20% of students of minority backgrounds and 8% of them receiving free and reduced lunches. In subsequent years, we will recruit teachers on a national basis. Teachers with access to technology and who serve underrepresented students will be given priority for participation. In Years 1 and 2, for each of the two modules, 120 students will participate in the feasibility studies taught by a teacher in MA and a teacher in CT (Table 3). In Year 3, a total number of 700 students will participate in the design studies, taught by 8 teachers from multiple states. In the last Year 4, 1,000 students from 10 teachers will participate in the pilot study. Some students may complete both modules.

Tables 1 and 2. Scoring Rubrics for the Explanation and Uncertainty Rationale Items, Respectively

Score/Descriptor	**Criteria**
Score 0: Blank/Off-task	Blank or irrelevant responses
Score 1: Irrelevant	Elicited non-normative ideas; Restated the claim selected or expressed; Incorrectly described the data
Score 2: Relevant knowledge/evidence	Correctly elicited one or more ideas listed above
Score 3: Single warrant	Correctly included two ideas mentioned above plus one of the links between the ideas
Score 4: Two or more warrants	Correctly included two or more links between valid idea

Score	**Uncertainty Source**	**Characteristics of Students' Rationale**
(Score 0)	• No information	No response; Irrelevant response; Restate the question
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
	• Specific knowledge/ability	Did not know specific scientific knowledge needed in the item set
	• 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
	• Specific 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

Power analyses. Statistical power analyses were conducted using the software program G*Power3 (Faul, Erdfelder, Buchner, & Lang, 2009). When conducting the power analysis, we considered different levels of effect sizes, as for a target power level, the required sample size changes with the desired effect size. Previous research on feedback has shown moderate to large effect sizes (0.40SD to 0.80SD) in improving learning (e.g., Azevedo & Bernard, 1995; Shute, 2008). To be conservative, we chose 0.25 as our target effect size (i.e., the lower the effect size, the greater number of students will be needed to achieve a certain power level). Our design studies in Year 3 (see the Research Question 2 & Table 3) will involve two experimental (i.e., automated scores with difference types of feedback) and one control conditions (i.e., automated scores without feedback). The design studies will achieve a statistical power of .95 at effect size of 0.25 with 700 students (350 for each module). For our pilot study in Year 4 (also see Research Question 2 & Table 3), data from 1,000 students, 500 for each module, will be used to study whether the effective feedback identified from the design studies should be offered with or without automated scores. The statistical power would be around .99 also for an effect size of 0.25.

Teacher Professional Development
Our researchers will pay frequent visits to schools during the week when the earth science instruction and assessment take place. Teachers will be able to review class-level performance on both embedded and posttest assessments through our synthesized reporting system. After-class meetings will be held with teachers of the same school to reflect on challenges with the technology, student reactions, and instructional modifications informed by the automated scores and feedback. The partner-teacher approach has been proven to be effective in improving teacher implementation of curriculum and assessment (Liu et al., 2011). Teachers will also decide the added value of automated feedback on items of different format. During the feasibility studies we will collaborate with teachers to create and refine feedback, considering student characteristics and relevance to instructional materials. We will work with teachers to investigate how information about student performance can be translated into effective science pedagogy.

During the feasibility studies, we will work with teachers to identify and overcome possible challenges in implementing automated scores and feedback. We will also work with teachers on how to use the diagnostics for more effective instruction. For the design and pilot studies we will continue to provide professional development through summer workshops and online tools. Teachers will participate in the workshops in Years 2 and 3 to understand the strengths and limitations automated scores and how they can take advantage of automated scores and immediate feedback for improved instruction. For example, teachers currently lack timely access to students' performance (Kerr et al, 2006; Shepard, 2000) and cannot make instructional decisions that are data driven. Automated scoring technology offers teachers the opportunities to use data to prioritize instructional activities. We will work with teachers to compare automated scores and teacher scores and discuss how automated scores may free teachers from the labor of scoring so that they can focus on instructional improvements.

Automated Scoring
ESAAF will develop automated scores using both *c-rater* and *c-rater-ML*. Both are automated scoring engines developed at ETS for scoring short-text items. *c-rater* has been supported by abundant research evidence, and has been used for many important assessments, including the NAEP Science Interactive Computer Tasks (e.g., Leacock & Chodorow, 2003; Sukkarieh & Pulman, 2005; Sukkarieh, Pulman, & Raikes, 2003; Sukkarieh, 2010). *c-rater* evaluates open-ended responses based on a set of clear, distinct concepts (Sukkarieh & Blackmore, 2009). The accuracy of *c-rater* scores depends on the cognitive skills and linguistic complexity of the responses. *c-rater* scoring involves four major steps: (a) ***model building**,* by which researchers identify one or more model responses which contain key concepts for the item; (b) ***natural language processing** (NLP)*, by which student and model responses are analyzed for linguistic features using NLP skills or knowledge representation skills; (c) ***main points identification***, by which the linguistic features are used to determine the absence or presence of a key concept in the student responses; and (d) ***scoring***, by which a score is assigned to a response based on the specified scoring rules. See CLASS in Prior Support for an example of *c-rater* scoring. In addition to an overall score, *c-rater* also provides scores for each concept, which can be used in feedback to students.

While *c-rater* scores open-ended responses deductively based on input from pre-assigned structures of concepts, *c-rater-ML* scores inductively by learning patterns in constructed-response item responses. *c-rater-ML* is based on modern machine learning technology and partially alleviates *c-rater's* need for response-specific work. *c-rater-ML* is based on a design in which numerous individual predictors of scores or features are generated, usually using supervised machine learning capabilities. As a relatively recent development, *c-rater-ML* was ranked top five among 140 international teams in the 2012 Hewlett Foundation's Short Answer Scoring Competition. For the purpose of this proposal, we pilot tested 1,027 student responses to the sample item in Fig 3 using *c-rater-ML*. The quadratic-weighted kappa was .80 for the explanation part and .70 for the uncertainty rationale part, indicating high agreement between human and machine scores (Williamson et al., 2012). The pilot results suggest that the earth science assessments we will be working with have great potential for automated scoring. The *c-rater* and *c-rater-*

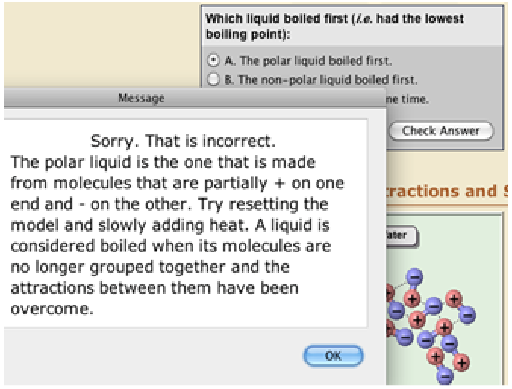
ML platforms will be integrated with Concord Consortium's existing curriculum and assessment platform for providing immediate scores and feedback.

Automated Scoring and Feedback System

In the proposed project, an interactive assessment and score reporting system will be created to support the automated scoring and feedback. It will be developed as a platform that can be also used for other science content areas beyond the topics investigated in this study. We will capitalize on the Concord Consortium's existing web portal and score reporting system developed by for the NSF-funded HAS and HAS:ESS projects (see Prior Support). We will integrate the Concord Consortium's portal system with the ETS automated scoring systems (*c-rater*, *c-rater-ML*) for constructed-response items, develop the capability to deliver instant automated scores to students, provide automated feedback on both multiple-choice and constructed-response items, and create a class-level reporting system for the teachers. Concord Consortium's current portal system includes an authoring system for developing online curriculum materials, interactive computational models, embedded assessments, and summary reports. It also provides class trajectories of student work and identified data for research uses.

Customized feedback. We will enhance the auto-feedback function to include constructed-response items. We will incorporate technological tools developed in the HAS project for feedback (see Prior Support). Currently those tools provide feedback to students at point of use. For example, students click a "check answer" button on multiple-choice items when they submit their answers (Fig 4). A message box pops up with feedback developed by the activity author. The feedback typically includes instruction for revisiting models, and/or hints as well as acknowledgement of correct responses. We will incorporate these features for constructed-response items as well as adding concept-based feedback, capitalizing on *c-rater*'s concept-based scoring.

Figs 4 & 5. Exemplar feedback to a MC response (L) and snapshot of student activity completion (R)



Which liquid boiled first (*i.e.* had the lowest boiling point):

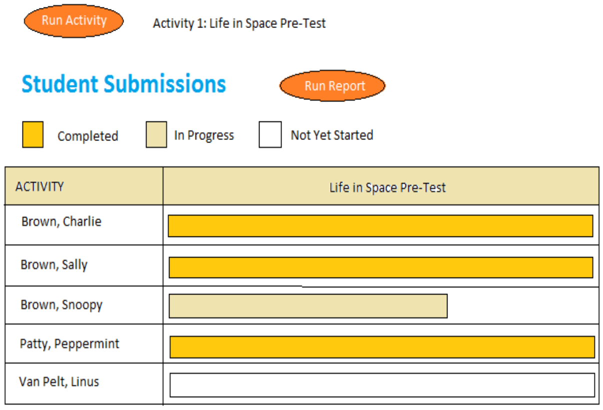
● A. The polar liquid boiled first.

○ B. The non-polar liquid boiled first.

ne time.

Message

Sorry. That is incorrect.
The polar liquid is the one that is made from molecules that are partially + on one end and - on the other. Try resetting the model and slowly adding heat. A liquid is considered boiled when its molecules are no longer grouped together and the attractions between them have been overcome.



Activity 1: Life in Space Pre-Test

Student Submissions

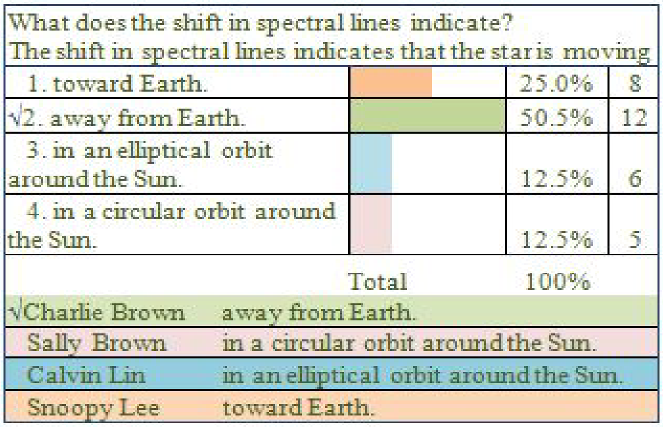
🟨 Completed ▨ In Progress □ Not Yet Started

ACTIVITY	Life in Space Pre-Test
Brown, Charlie	🟨
Brown, Sally	🟨
Brown, Snoopy	▨
Patty, Peppermint	🟨
Van Pelt, Linus	□

Class-level snapshot. We will also improve Concord Consortium's current reporting system to provide both at-a-glance indicators of students' performance and document learning progressions. Currently teachers can have real-time information on student completion of any given activity (Fig 4). Teachers also have instant access to class-level performance distributions on MC items (Fig 5). We will extend the system to including constructed-response items, and will create a class-level snapshot at both the item and the assessment level (Fig 6).

Progress monitor. We will develop methods for displaying the progresses students made through the activities, which will visually represent changes in student performance from pre, embedded, to posttest on the same items. Additionally, we will monitor students' interaction with the assessment system and track their answer-changing behavior as result of feedback.

Fig 6. Class-level performance on an MC item



What does the shift in spectral lines indicate? The shift in spectral lines indicates that the star is moving			
1. toward Earth.	🟧	25.0%	8
√2. away from Earth.	🟩	50.5%	12
3. in an elliptical orbit around the Sun.	🟦	12.5%	6
4. in a circular orbit around the Sun.	🟥	12.5%	5
	Total	100%	
√Charlie Brown	away from Earth.		
Sally Brown	in a circular orbit around the Sun.		
Calvin Lin	in an elliptical orbit around the Sun.		
Snoopy Lee	toward Earth.		

The system will also document the frequency of each feedback delivery. All this information can be of great value for teachers as they identify areas in which students are having difficulties. Teachers can then revise instruction to effectively address problematic concepts. We will draw on the infrastructure developed for the NSF-supported HAS projects (see Prior Support). In addition, to cater to the increasing variety of operating systems, and we will enhance technological capabilities to deliver materials for computers, tablets, and other operating systems.

Research Questions (RQ)

RQ 1: To what extent can automated scoring tools such as c-rater and c-rater-ML, diagnose students' explanations and uncertainty articulations as compared to human diagnosis?

In Years 1 and 2, we will develop and evaluate the accuracy of automated scoring for the climate change and fresh water availability modules, respectively. Specifically, we will apply the automated scoring technology to students' responses (*n* >1,000) to the constructed-response explanation and uncertainty rationale items created and validated in the NSF-funded HAS and HAS:ESS projects.

We will examine how well scores generated by automated scoring methods agree with human scores. Multiple criteria will be used to evaluate the quality of automated scores, including the (a) quadratic-weighted kappa, (b) Pearson correlation, (c) degradation of the human/machine score agreement from the human/human score agreement, and (d) standardized mean score difference between the machine scores and human scores. The Kappa coefficient indicates the proportion of agreement beyond that expected by chance and is scaled to range from -1 to 1, with -1 indicating poorer than chance agreement, 0 pure chance agreement, and 1 perfect agreement. We plan to adopt the Landis and Koch (1977) rules for the strength of agreement for the Kappa coefficient: poor (≤.00), slight (.00 -.20), fair (.21-.40), moderate (.41-.60), good (.61-.80), and very good (.81-1.00). The degradation value represents the difference in agreement between human/human and human/automated scores. The introduction of degradation recognizes the dependence of the performance of automated scores on human scoring agreement. The degradation of automated scoring agreement from human agreement should not be more than .10 as a guideline for operational practice (Williamson et al., 2012). The standardized mean score difference between the automated and human scores is another criterion indicating the performance of automated scores. Standardization ensures that the two sets of scores are compared on similar scales, and should not be more than .15 (Williamson et al., 2012).

After the automated scoring models are established, we will implement the automated scoring in classrooms as part of the feasibility studies in Years 1 and 2 for the two modules, respectively. For each module, we will work with two teachers and approximately 120 students to implement the week-long curriculum and test the feasibility of automated scoring. We will randomly split students within a teacher into the automated scoring and immediate feedback condition, and teacher score and delayed feedback condition. We will use independent sample *t*-tests to compare students' review and revision behavior and also score gains as a result of revision after receiving feedback. These studies will provide evidence of feasibility of implementation by showing that teachers can implement the modules with the automated scoring in a classroom setting.

RQ 2: How should feedback be designed and delivered to help students improve scientific argumentation? How does students' use of feedback relate to their learning progression during instruction and learning outcome at the end of instruction? Should feedback be offered with or without scores?

Literature shows that the effects of feedback are significantly determined by the feedback type and delivery method (Anderson, Magill, & Sekiya, 2001; Kulik & Kulik, 1988; Shute, 2008). In order to determine the optimal type of feedback that can maximize student learning, we will conduct two design studies in Year 3 to investigate how students use different types of automated feedback during the instruction, and what learning outcomes emerge from these varied uses. Each study will have two feedback conditions, and a control condition to which students within a teacher will be *randomly* assigned. Students in the control condition will learn the module without feedback, but will be told that they can review and revise their answers during assessment. To eliminate the possible negative impact of students receiving lower scores, students will only see the feedback without the scores, but note that the automated scoring technology is required for the offering of the feedback. In Year 4, we will conduct a pilot study to implement the automated feedback proven to be effective based on the results from Year 3, and test whether feedback should or should not be offered with autoscores.

Design Study 1: Automated content feedback vs. automated epistemic feedback. Open-ended writing is required in explanations and uncertainty rationale in our argumentation items. As in any scientific writing, student writing for explanations and uncertainty rationale can be conceived to occur in content and rhetorical spaces (Bereiter & Scardamalia, 1987). Writing in the content space refers to theoretical and empirical bases that validate the argument, while writing in the rhetorical space refers to the structure and the requirements related to an effective argument. In the literature, the distinction in supports built for content and for rhetorical spaces is sometimes made as generic (content-independent) versus specific (content-dependent). In argumentation research, writing in the rhetorical space is often referred to as epistemic practice (Sandoval, 2003). The literature debates about which type of scaffolds is more effective in improving students' open-ended explanations (Butcher & Kintsch, 2001; Davis, 1998). As a first design study, we will investigate how automated content and epistemic feedback are used by students, how they help students formulate arguments, and which feedback is more effective.

- Automated content feedback will provide content-related diagnostic information. Example: "You recognized the importance of the reflection of sunlight by the Earth's surface (albedo), but did not elaborate how albedo affects temperature."
- Automated epistemic feedback will provide rhetorical diagnostic information about students' level of explanations and uncertainty rationale such as "In your explanation, you restated your claim but did not include data," or "In explaining your uncertainty, you mentioned about your ability or knowledge, but did not mention uncertainty related to models."

Design Study 2**: **Automated diagnostic feedback vs. automated suggestive + diagnostic feedback. After the design study 1, we will determine whether to focus on content or epistemic feedback. In the design study 2, we will determine whether the content or epistemic feedback should be given in a form of diagnostic or diagnostic plus suggestive feedback. Diagnostic feedback contains information regarding students' ability to explain the claim and articulate sources of uncertainty. Suggestive feedback, in addition to the diagnosis of student performance, also points students to instructional steps in the climate change and fresh water availability modules that are relevant to the assessment task at hand. The effectiveness of suggestive feedback is reported in some existing studies (Black & Wiliam, 1998; Narciss & Huth, 2004), but it is unknown if the conclusion remains true when the automated feedback is involved.

- Automated diagnostic feedback will show students their level of explanations and uncertainty rationale according to the scoring rubrics described in the Curricula, Assessments and Rubrics. Rubrics for explanations and uncertainty rationale have been tested in a series of large scale scientific argumentation assessment trials with a reliability of .91.

- Automated diagnostic+suggestive feedback will show students their level of explanation and uncertainty rationale, as well as where in the climate change or fresh water modules they can get more information. Example: "You recognized the reflection of sunlight by the Earth's surface is important in determining global temperature changes but did not explain how. To know more about this, go back to this activity step (hyperlinked text to the related activity step)."

In Year 3, a total of eight teachers and their students (approximately *n*=700) will participate in the design studies. We will administer the identical pre/posttests and embedded assessments to students. Automated feedback will be provided for both embedded and posttests but not for pretests, as the pretest scores will be used as a control for prior learning. We will compare the two feedback and control groups on (a) review and revision behavior (e.g., frequency), (b) revision outcomes on both the embedded and posttest, and (c) pre/posttest gains using both initial and revised posttest scores. Since each design study involves two feedback and one control groups, we will use an ANOVA to investigate any possible differences in the outcomes described above. We will analyze the interaction between student science ability as measured by the pretest and the feedback effects. In addition, we will interview a small group of students in the feedback conditions to understand their experiences with the automated feedback, and their perceived usefulness of the feedback. We will analyze students' actions after receiving feedback (e.g., going back to the instructional steps as suggested by the feedback). We will study students' review and revision patterns on multiple-choice and constructed-response items to see if there is any differential effect of feedback. Thusly, our automated feedback will be continuously refined through the results of the design studies.

Pilot Study: Effective feedback with or without automated scores. The feedback identified to be effective from the above-mentioned design studies will be implemented at a larger-scale among students in Year 4. In addition, as literature suggests that students may be discouraged by grades or scores (Butler, 1987; Shute, 2008), we will test the effect of automated feedback with or without automated scores or grades on individual items. A total of 10 teachers and their students (*n*=1,000) will participate in the pilot study. Students within a teacher will be randomly assigned to the two feedback conditions: feedback only or feedback plus scores. We will compare their review and revision behavior using independent sample *t*-tests. We will also compare students' improvements on embedded assessments and pre/post gains after they receive automated feedback with or without scores. We plan to use ANCOVA to investigate student learning progressions with posttest scores as the outcome variable, feedback condition as a fixed factor, and various covariates including pretest scores, gender, language, and grade.

RQ 3. How do teachers use automated scores and feedback to improve their instructional practices? How do teachers' conceptions change about automated scores and assessment?

We will adopt both qualitative and quantitative approaches to address the research questions. We will conduct a survey asking about teachers' perceptions of automated scores and feedback both before and after the implementation of the curriculum and assessments in the feasibility, design, and pilot studies. We will analyze the differences in teachers' perceptions. We will conduct classroom observations during the feasibility and design studies to understand how teachers use the class-level automated feedback to modify and improve instruction. For example, if the autoscores and feedback suggest that a large portion of the students demonstrate misunderstanding of a certain concept, will the teacher take immediate actions to revisit the instructional step relevant to that concept? We will interview teachers to learn about how feedback interacts with teachers' pedagogical practices to produce better learning. We will also elicit teachers' input and opinions on how to design professional development resources to help them use the technologies. In professional development workshops in Years 2 and 3, we will provide ample opportunities for teachers to understand automated scoring, help design effective feedback, and exchange ideas about how to use the information at the class-level. We will also help teachers understand how to assess argumentation embedded in the science topics we chose. We will link the observation and teacher survey data with student learning outcomes to investigate how teachers' different uses of automated scoring and feedback are associated with student learning outcomes at the end of each module.

In addition, the interactive system ESAAF will give teachers access to both class-level and individual performance records. Teachers will also be able to share insights, student progress, and implementation plans with fellow teachers, and discuss challenges and experiences. We will also elicit teacher opinions on the most valuable features about the score reporting system. One of the great challenges of effective teaching is teachers' lack of real-time student data (Sisk-Hilton, 2009). Automated scoring and feedback relax the constraint so that the assessment, instruction, and learning processes can be streamlined for improved efficiency.

EVALUATION AND EXTERNAL REVIEW

This project plans to be ably reviewed, advised, and evaluated by the leading experts and practitioners in the field. These include **Charles Anderson,** (Professor, Michigan State University), **Mark Shermis** (Professor, University of Akron), **Ayita Ruiz-Primo** (Associate Professor and Director, University of Colorado, Denver), **Xiufeng Liu** (Professor, State University of New York, Buffalo), and **Rick Dees** (Science teacher of 19 years, Huntley Project High School). The research team will elicit advisor expertise in automated scoring (Shermis), feedback (Anderson, Ruiz-Primo, Dees), psychometrics (Liu, Ruiz-Primo), earth science content (Anderson, Dees), and assessment (Anderson, Liu, Ruiz-Primo).

We intend to have two advisory panel meetings each year, one through videoconference and one face-to-face meeting to discuss updates, evaluate progress, and plan for next steps. **Xiufeng Liu** will assume the role of **external evaluator**. Dr. Liu has extensive experiences in science assessment, science instruction, and measurement, and is currently co-editor of the *Journal of Research in Science Teaching*. Dr. Liu will independently monitor project progress compared to the timeline specified in the proposal, evaluate the quality of assessments from both psychometric and content perspectives, review the scoring rubrics, and review the psychometric analysis. Dr. Liu is also anticipated to evaluate the reports, presentations, and materials produced from this project. The evaluator will participate in monthly management and evaluation meetings for project updates, and provide feedback on how to effectively execute the research goals. The evaluator will also participate in annual face-to-face project meetings, and work with the advisory panel members in providing recommendations to the research team regarding project progress.

DISSEMINATION

This project is intended to create a rich legacy of assessments, rubrics, feedback, and professional development materials. We will proactively disseminate the research findings from this project. We will use multiple methods to reach a wide audience, including a website, presentations, research reports, journal articles, Wiki pages, and Facebook groups. The website will feature free assessment items, rubrics for both human and automated scoring, and links to project papers, reports, and presentations. The assessments, rubrics, design principles, and technologies will be *open source* through the ETS and CC's project websites. To reach multiple stakeholders, we will make presentations at national conferences such as AAAS, AERA, NCME, ICLS, and NARST. We also plan to participate in policy forums and author articles for high-circulation journals such as *Science*, *Educational Researcher, Journal of Research in Science Teaching,* and *Journal of the Learning Sciences.*

BROADER IMPACTS

The automated scoring and feedback system developed in ESAAF has great potential to transform science teaching and learning through formative assessments at multiple levels. At a proximal level, the online curriculum modules used in ESAAF will be empowered by automated scoring and quality feedback technologies and will be made available for free to all future learners, teachers, and researchers beyond the research participants outlined in this proposal. In 2013 alone, over 16,000 students taught by 75 teachers have used our curricula without explicit support of the project staff, and the numbers are expected to increase significantly when new features of automated scores and feedback are incorporated with purposeful dissemination and support. The effective feedback identified from this project can also be used by teachers with or without the automated scores. At a macro level, the automated scoring and feedback approach provide a model on how formative assessments can be integrated into learning opportunities. Such integration is currently hindered in many classrooms due to the lack of timely feedback on formative, constructed-response assessments. At the national level, the goals of the ESAAF project are closely aligned with those of the *Race to the Top* initiative in that both the *Smarter Balanced* and the *PARCC* consortia are exploring how to take advantage of automated scoring to increase the use and value of formative assessment.

Table 3. ESAAF Research and Development Timeline

YEAR 1	YEAR 2	YEAR 3	YEAR 4
AUTOMATED SCORING (AS)			
Develop AS for 11 explanation and 11 uncertainty rationale items for climate change	Develop AS for 11 explanation and 11 uncertainty rationale items for fresh water	Refine AS for both climate change and fresh water modules	Finalize AS for both climate and fresh water modules
FEEDBACK			
Design feedback for 11 explanation and 11 uncertainty rationale items for climate change	Design feedback for 11 explanation and 11 uncertainty rationale items for fresh water	Refine feedback for both climate change and fresh water modules	Finalize feedback for both climate and fresh water modules
RESEARCH			
Feasibility study to see if AS and feedback work for climate change; Evaluate automated vs. teacher feedback; Teacher (*n*=2); Student (*n*=120).	***Feasibility study*** to see if AS and feedback work for fresh water; Evaluate automated vs. teacher feedback; Teacher (*n*=2); Student (*n*=120).	***Design studies*** to identify effective feedback through random assignment (i.e. content vs. epistemic and diagnostic vs. diagnostic+suggestive); Teacher (*n*=8; 4 for each module); Student (*n*=700).	***Pilot study*** to see if effective feedback should be offered with or without scores; Teacher (*n*=10; 5 for each module); Student (*n*=1,000).
PROFESSIONAL DEVELOPMENT			
Elicit teacher input in feedback design	Summer workshop for the 8 teachers who will participate in the design studies, to familiarize them with automated scoring and feedback	Summer workshop with the 10 teachers who will participate in the pilot study in Year 4	Ongoing support of teachers' use of automated scores and feedback in formative assessment
TECHNOLOGY INNOVATIONS			
Develop platform; Integrate ETS autoscoring & CC assessment system	Incorporate auto feedback into platform; Develop class-level snapshot for teacher review	Improve platform's delivery of scores and feedback for both students and teachers	Refine platform features on review and revision

EXPERTISE

ESAAF collaborators include **Educational Testing Service (ETS), Concord Consortium (CC)**, and **University of California, Santa Cruz (UCSC)**. ETS, a non-profit educational organization and a world-class leader in assessment, automated scoring, and measurement research, is the lead institution, and will be responsible for the development of automated scoring, research design, quantitative analyses, financial oversight, and overall project performance. Amy Pallant at CC will be the Co-PI responsible for technology development, assessment review, and professional development. Hee-Sun Lee at UCSC is the Co-PI responsible for assessment and feedback development, participating in research studies, and coordinating with ETS and CC on professional development. The diverse expertise and rich experiences our collaborators bring to ESAAF are expected to significantly contribute to the successful implementation of this project. ETS plans to actively coordinate with the partners to ensure the project meets its goals and objectives on time. We intend to have bi-weekly teleconferences among key partners to report updates, discuss potential challenges, and plan for next steps. In the meetings we plan to use videos to display technological systems if needed and use Google Drive to share materials. ETS will have weekly internal staff meetings gathering updates on milestones and deliverables. In addition to these meetings and the advisory board meeting mentioned above, we've budgeted professional development workshops in the summers in Years 2 and 3 to provide training to teachers on the use of automated scores and feedback and discuss how the diagnostic information can be used to improve instruction.

Ou Lydia Liu, **ETS; PI.** Dr. Liu is Managing Senior Research Scientist at ETS. She will oversee the research design, automated scoring, and quantitative analyses, and be responsible for the overall progress of the project. Dr. Liu is currently Co-PI on three NSF-funded, five-year projects concerning innovative science assessment and automated scoring. She has published more than 30 articles in top measurement and science education journals. She received the *ETS Presidential Award* for outstanding research in 2008, and the *Jason Millman Promising Measurement Scholar Award* from the National Council on Measurement in Education in 2011. Dr. Liu holds a PhD in Quantitative Methods and Evaluation at the University of California, Berkeley.

Amy Pallant, **CC; Co-PI**. She will be responsible for technology development, feedback development, assessment review, and professional development of the project. She will direct the development of the synthesized assessment and scoring report system, and coordinate the technology development with ETS. She is the PI for the NSF-funded HAS and HAS:ESS projects (DRL-1220756, DRL-0929774). She has been the project manager, educational researcher, and curriculum developer on the award winning Molecular Workbench projects. Amy holds an M.A. in Science Education from Harvard and a B.A. in Geology from Oberlin College.

Hee-Sun Lee; **UCSC; Co-PI**. Dr. Lee will be responsible for assessment and feedback development, participating in research studies, and coordinating with ETS and CC on professional development. Dr. Lee is currently an Adjunct Associate Professor at the UCSC. She specializes in inquiry-based curriculum and assessment development and evaluations of innovative curriculum materials. She is Co-PI on an NSF-funded project (DRL-1220756) on the HAS:ESS, and directed a large-scale NSF-funded assessment research program at the Technology-Enhanced Learning in Science (TELS) Center. She earned a Ph.D. in Science Education from the University of Michigan.

Katrina Crotts Roohr; ETS; Quantitative Scientist/Psychometrician. Dr. Roohr is an Associate Research Scientist at ETS. She will assist in the research design and quantitative and psychometric analyses for this study. Dr. Roohr holds an Ed.D. in Psychometric Methods, Educational Statistics, and Research Methods from the University of Massachusetts Amherst. Prior to her graduate studies, she received a B.A. degree (with honors) in Psychology with minors in Mathematics and Spanish from Westfield State University.

Mike Heilman, ETS; Automated Scoring Scientist. Dr. Heilman's research includes various applications of natural language processing and machine learning to educational problems, including the analysis of argumentation, the detection of grammatical errors, and the scoring short answer questions. He is currently the chief automated scoring scientist for the CLASS project (see Prior Support). He received a Ph.D. in Language and Information Technologies from Carnegie Mellon University. Prior to his graduate studies, he received a B.S. degree (with honors) in computer science and a B.A. degree (with honors) in Japanese language, both from the University of Notre Dame.

John Blackmore; ETS; Automated Scoring Engineer. Blackmore is Lead Software Developer at ETS. He will be responsible for rubric review for automated scoring, building automated scoring models, and statistical analysis of the agreement between human and machine scores. He is the lead automated scoring engineer for the NSF-funded CLASS project (DRL-1119670).

EARTH SCIENCE ASSESSMENTS WITH AUTOMATED FEEDBACK (ESAAF)

References

Allchin, D. (2012). Teaching the nature of science through scientific errors. *Science Education, 96*(5), 904-926.

Anderson, D., Magill, R. A., & Sekiya, H. (2001). Motor learning as a function of KR schedule and characteristics of task intrinsic feedback. *Journal of Motor Behavior, 33,* 59-66.

Attali, Y., & Powers, D. (2008*). Effect of immediate feedback and revision on psychometric properties of open-ended GRE subject test items* (ETS GRE Board Research Report No. 04-05; ETS RR-08-21). Princeton, NJ: Educational Testing Service.

Attali, Y., Powers, D., Freedman, M., Harrison, M., & Obtez, S. (2008). *Automated scoring of short-answer open-ended GRE subject test items* (ETS GRE Board Research Report No. 04-02; ETS RR-08-20). Princeton, NJ: Educational Testing Service.

Aufschnaiter, C. V., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching, 45*, 101-131.

Azevedo, R., & Brenard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13,* 11-127.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*, 26-55.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Berkshire, England: McGraw-Hill Education.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7–74.

Bricker, A., & Bell, P. (2008). Conceptualizations of argumentation from science studies and the learning sciences and their Implications for the practices fo science education. *Science Education, 92*, 473-493.

Brosvic, G. M., & Cohen, B. D. (1988). The horizontal vertical illusion and knowledge of results. *Perceptual and Motor Skills, 67,* 463–469.

Brown, G. T. L., Harris, L. R., & Harnett, J. (2012). Teacher beliefs about feedback within an assessment for learning environment: Endorsement of improved learning over student well-being. *Teaching and Teacher Education, 28,* 968-978.

Burstein, J., & Marcu, D. (2003). Automated evaluation of discourse structure in student essays. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 209-229). Mahwah, NJ: Routledge.

Butcher, K. R., & Kintsch, W. (2001). Support of content and rhetorical processes of writing: Effects on the writing process and the written product. *Cognition and Instruction, 19*(3), 277-322.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79*(4), 474-482.

Chin, C., & Osborne, J. (2010). Students' questions and discursive interaction: Their impact on argumentation during collaborative group discussions in science. *Journal of Research in Science Teaching, 47*(7), 883-908.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*, 175-218.

Clark, D., Sampson, V., Weinberger, A., & Erkens, G. (2007). Analytic frameworks for assessing dialogic argumentation in online learning environments. *Educational Psychology Review, 19*, 343-374.

Clarke, S. (2003). *Enriching feedback in the primary classroom*. London, England: Hodder and Stoughton.

Corbett, A. T., & Anderson, J. R. (1989). Feedback timing and student control in the LISP intelligent tutoring system. In D. Bierman, J. Brueker, & J. Sandberg (Eds.), *Proceedings of the Fourth International Conference on Artificial Intelligence and Education* (pp. 64–72). Amsterdam, The Netherlands: IOS Press.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438-481.

Davis, E. A. (2003). Prompting middle school science students for productive reflection: Generic and directed prompts. *The Journal of the Learning Sciences, 12*(1), 91-142.

Dihoff, R. E., Brosvic, G. M., & Epstein, M. L. (2003). The role of feedback during academic testing: The delay retention effect revisited. *The Psychological Record, 53,* 533–548.

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education*, *38*, 39-72.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*, 915-933.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1,* 3-31.

Graesser, A. C. (2011). Learning, thinking, and emoting with discourse technologies. *American Psychologist, 66*(8), 746-757.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Onley, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 180-92.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language, 25*, 282-306.

Institute of Education Sciences & National Science Foundation. (2013). *Common guidelines for education research and development.* Washington, DC: U.S. Department of Education.

Intergovernmental Panel on Climate Change. (2007). Climate change 2007: Synthesis report: Summary for policymakers. Retrieved on Jan 2, 2013 from http://www.ipcc.ch/pdf/assessment-report/ar4/syr/ar4_syr.pdf

Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, R. A. (1999). "Doing the lesson" or "doing science": Argument in high school genetics. *Science Education, 84*, 757-792.

Kerr, M. S., Rynearson, K., & Kerr, M. C. (2006). Student characteristics for online learning success. *The Internet and Higher Education, 9,* 91-105.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284.

Koedinger, K. R., McLaughlin, E. A., & Heffernan, N. T. (2010). A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research, 43*(4), 489-510.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development, 74*(5), 1245-1260.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58,* 79–97.

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

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice, 23*(3), 6-14.

Leacock, C., & Chodorow, M. (2003). Automated grammatical error detection. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 186-199). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ledley, T. S., Dahlman, L., McAuliffe, C., Haddad, N., Taber, M. R., Domenico, B., Lynds, S., & Grogan, M. (2011). Making earth science data accessible and usable in education. *Science, 333*(6051), 1838-1839.

Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24,* 115-136.

Lee, H. -S., Liu, O. L., Pallant, A., Pryputniewicz, S., Crotts, K., & Buck, Z. (in press). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching*.

Lee, H. –S., Pallant, A., Pryputniewicz, S., & Liu, O. L. (2013, April). *Measuring students' scientific argumentation associated with uncertain current science*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Rio Grande, Puerto Rico.

Lee, H.-S., & Songer, N. B. (2003). Making authentic science accessible to students. *International Journal of Science Education, 25*(8), 923-948.

Linn, M., Liu, O.L., Ryoo, K., & Madhok, J. (2012, April). *Teaching and assessing scientific thinking: Online inquiry units with automated scoring.* Paper presented at the 2012 Annual Conference of the American Educational Research Association, Vancouver, BC.

Liu, O.L., Brew, C., Blackmore, J., Gerard, L., & Madhok, J. (2013, April). *Automated scoring for inquiry science assessment: An application of c-rater.* Paper presented at the 2013 Annual Conference of the National Council on Measurement in Education, San Francisco, CA.

Liu, O. L., Lee, H.-S., & Linn, M. C. (2011). Measuring knowledge integration: A four year study. *Journal of Research in Science Teaching, 48,* 1079-1107.

Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction, 22*, 219-290.

Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerised marking of free-text responses. P*roceedings of the 6th CAA International Computer Assisted Assessment Conference, Loughborough, United Kingdom*.

Molden, D. (Ed.). (2007). *Water for food, water for life: A comprehensive assessment of water management in agriculture*. London, UK: International Water Management Institute.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multimedia learning. In H. M. Niegemann, D. Leutner, & R. Brunken (Ed.), *Instructional design for multimedia learning* (pp. 181–195). Munster, Germany: Waxmann.

National Council of Teachers of English. (2008). *Statement on class size and teacher workload: Secondary.* Retrieved from http://www.ncte.org/positions/statements/classsizesecondary

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

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

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994-1020.

Pallant, A., Pryputniewicz, S., & Lee, H. –S. (2012). Exploring the unknown: Fostering critical thinking in earth and space science. *The Science Teacher, 79*(3), 60-66.

Pallant, A., & Tinker, R. (2004). Reasoning with atomic-scale molecular dynamic models. *Journal of Science Education and Technology, 13*, 51-66.

Partnership for Assessment of Readiness for College and Careers. (2012). *PARCC progress updated: March 2012.* Retrieved from PARCC website http://www.parcconline.org/sites/parcc/files/PARCC%20Progress%20Report%20-%20FINAL.pdf

Quellmalz*,* E.*,* Timms*,* M., & Buckley*,* B. *(*2005*). Using science simulations to support powerful formative assessments of complex science learning*. San Francisco, CA: WestEd.

Ritter, K. (2012). *UN climate scientist: Sandy no coincidence.* USA Today. Retrieved from http://www.usatoday.com/story/weather/2012/11/27/climate-change-hurricane-sandy/1730251/

Rogers, P. (2008). Coping with global warming and climate change. *Journal of Water Resources Planning and Management*, *134*(3), 203–204.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy, and Practice, 5,* 77-84.

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35,* 535-550.

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, *92*, 447-472.

Sandene, B., Horkay, N., Bennett, R. E., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project, research and development series* (NCES 2005-457). Washington, DC: U.S. Department of Education, National Center for Educational Statistics.

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *The Journal of the Learning Sciences, 12*(1), 5-51.

Shepard, L. A. (2000). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice, 28*(3), 32-37.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78,* 153-189.

Singer, S. R., Hilton, M. L., & Schweingruber, H. A. (2006). *America's Lab Report: Investigations in High School Science*. Washington, National Academies Press.

Sisk-Hilton, S. (2009). *Teaching and learning in public: Professional development through shared inquiry*. New York, NY: Teachers College Press.

Smarter Balanced Assessment Consortium. (2012). *Smarter Balanced Assessment Consortium: General item specifications.* Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/ItemSpecifications/GeneralItemSpecifications.pdf

Sukkarieh, J. Z. (2010, July). *Maximum entropy for the automatic content scoring of free-text responses*. Paper presented at MaxEnt 2010, Chamonix Mont-Blanc, France.

Sukkarieh, J. Z., & Blackmore, J. (2009). c-rater: Automatic content scoring for short constructed responses. In *Proceedings of the 22nd International FLAIRS Conference* (pp. 290-295). Sanibel Island, FL: Association for the Advancement of Artificial Intelligence.

Sukkarieh, J. Z., & Pulman, S. G. (2005). Automatic short answer marking. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP* (pp. 9-16). Stroudsburg, PA: Association for Computational Linguistics.

Sukkarieh, J. Z., Pulman, S. G., & Raikes, N. (2003, October). *Auto-marking: Using computational linguistics to score short, free text responses*. Paper presented at the 29th annual conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

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

VanLehn, K., Jordan, P., & Litman, D. (2007). Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. *Proceedings of SLaTE Workshop on Speech and Language Technology in Education*.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6,* 103–118.

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

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections, 21*. Retrieved from http://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf