# Understanding the Transformative Potential of Spoken Assessments of Science Understanding for Young Learners

Wayne Ward

Ron Cole

Brandon Helding

Finbarr Sloane (Arizona State University)

**Project Summary**

The goal of *Understanding the Transformative Potential of Spoken Assessments of Science Understanding for Young Learners* project is to investigate the transformative potential of evaluating young learners' understanding of science (and STEM content in general), through spoken assessments of science. We propose that spoken assessments of science knowledge will produce more engaging and accurate estimates of students' science learning and achievement, and will in fact produce higher estimates of young learners' science knowledge than written assessments currently used in 3rd through 5th grade assessments associated with the FOSS curriculum. Moreover, we argue that spoken assessments offer the promise of developing measures for our youngest (pre-k through second grade) students' science knowledge and abilities, through spoken dialogs with pedagogical agents in multimedia environments with strong arguments for validity and reliability. Specifically, we have two aims: 1) assess benefits of incorporating spoken prompts into written assessments, and 2) investigate spoken dialogs between a student and a virtual or human teacher testing if they can produce detailed and accurate assessments of science understanding as well as the ability to verbalize complete and accurate science explanations.

**INTELLECTUAL MERIT:** The spoken assessments we propose to develop and evaluate are designed to provide teachers and students with powerful tools that can be integrated into science curricula to meet the learning goals described in the New Science Framework. The New Science Framework is consistent with the conclusions and recommendations of the National Research Council, highlighting that "Students who are proficient in science: 1. know, use, and interpret scientific explanations of the natural world; 2. generate and evaluate scientific evidence and explanations; 3. understand the nature and development of scientific knowledge; and 4. participate productively in scientific practices and discourse". These principles speak to the role of discourse in building scientific knowledge. A key theme of our research is that it is efficient, effective, and *necessary* to provide spoken assessments of science knowledge to gain an accurate estimate of student understanding. Spoken assessments are an essential component of a paradigm shift that is required for educators and students to achieve one goals in the New Science Framework-- to help teachers learn how to facilitate conversations where students learn, exercise and become proficient in scientific discourse and argumentation. This can only happen if assessments of students' science knowledge evaluate these abilities. Currently, national and state assessments of students' science learning and achievement use written tests. Written questions and answer choices are also embedded in science curricula, such as the FOSS program (that will be used in our study). There is strong evidence to support that written tests greatly underestimate science learning by the majority of young learners in the U.S., who have difficulty reading and understand English text. We hypothesize that spoken assessments, including assessments that provide language-minority students with first language support, will be more engaging and effective instruments for understanding what students' have learned about science.

**BROADER IMPACTS:** To understand the transformative potential of spoken dialogs to improve our ability to measure understanding of science, consider how science learning and achievement is currently measured for kindergarten through elementary school (K-5) children in the U.S. today. All states, as mandated by federal law, administer tests of science knowledge to 3rd, 4th or 5th grade students, with the vast majority of states testing 5th grade students. All state tests include written questions and response alternatives, which require students to read and understand questions and answer choices. A student who cannot read or write well is likely to score poorly on these tests, regardless of science knowledge; this makes low-performing ELL students indistinguishable from other students that simply do not understand science. Similarly, there is compelling evidence that written assessments can greatly underestimate students' science learning and achievement. Spoken assessments will produce more engaging test-taking experiences, and result in more accurate estimates of students' science knowledge and abilities. This will address needs cited by the National Literacy Panel when they concluded that "language-minority students rarely approach the same levels of proficiency in text-level skills achieved by native English speakers" and NAEP results suggested, "the reading achievement gap between Hispanic and White 4th grade students" translated to Hispanic students achieving less than half White 4th grade students.

## 1. Significance

The goal of this project is to investigate the transformative potential of evaluating young learners' understanding of science (and STEM content in general), through spoken assessments of science. We propose that spoken assessments of science knowledge may produce more engaging and more accurate estimates of students' science learning and achievement, and will in fact produce higher estimates of young learners' science knowledge than written assessments currently used in elementary grades. Moreover, we will argue that spoken assessments could offer the promise of developing measures of our youngest (pre-k through second grade) students' science knowledge and abilities.

In order to understand the transformative potential of spoken dialogs to improve our ability to measure young learners' understanding of science, it is necessary to consider how science learning and achievement is currently measured for kindergarten through elementary school (K-5) children in the U.S. today. In elementary grades, children are administered written tests of science knowledge. All states, as mandated by federal law, administer tests of science knowledge to 3rd, 4th or 5th grade students, with the vast majority of states testing 5th grade students. All state tests include written questions and response alternatives, which require students to read and understand questions and answer choices. The National Assessment of Educational Progress (NAEP) administers written tests of science achievement to students in 4th, 8th and 12th grades. The NAEP [1] measured students' knowledge in physical science, life science, and earth and space science using a new science framework designed "to keep assessment content current with key developments in science, curriculum standards, assessments, and research." The results indicated that "Thirty-four percent of students at grade 4, some 30 percent of students at grade 8, and 21 percent of students at grade 12 performed at or above the Proficient level in the 2009 science assessment. One percent of 4th-grade students, 2 percent of 8th-grade students, and 1 percent of 12th-grade students performed at the Advanced level" [1].

A major problem with written assessments is that students who do not read proficiently will test poorly on written assessments of science knowledge. Put simply, a student who can read and write proficiently, and understands the science, is likely to score well on a written test. A student who reads and writes proficiently, but does not understand the science, is likely to score poorly on the same test. But a student who cannot read or write well is likely to score poorly on the same test, regardless of their science knowledge, because they cannot read nor understand the questions or answer choices, or produce answers that express their ideas well in writing. The result is that, based solely on the results of a written science test, we cannot accurately assess students' science knowledge and abilities. In essence, we have an unavoidable, psychometric confound to any validity and reliability arguments we pose about assessments of students' science knowledge.

There is compelling evidence that written assessments can greatly underestimate students' science learning and achievement for many students. Cromley [2] reported correlations of greater than .80 between individuals' reading scores and science scores on standardized tests of reading literacy and science literacy across countries; these tests were administered by the Program for International Student Assessment (PISA) a system of international assessments that focuses on 15-year-olds' capabilities in reading literacy, mathematics literacy, and science literacy [2]. In the U.S., about two thirds of U.S. students are not proficient readers; the 2009 NAEP [3] indicated that only 33% of 4th, 8th and 12th graders scored as Proficient or Advanced readers. Given that the majority of students in U.S. public schools do not achieve proficient or advanced reading levels on national tests, we can assume that written science assessments underestimate many students' science learning and achievement. Moreover, the report of the National Literacy Panel concluded that "language-minority students rarely approach the same levels of proficiency in text-level skills achieved by native English speakers" [4, p.10]. The 2009 NAEP indicated that "the reading achievement gap between Hispanic and White 4th grade students" [3] was 25 points; this translates to Hispanic students achieving less than half the performance of White 4th grade students and, we can infer ,would also transfer over to English-based, written science assessments.

Another problem with written tests is that they are often stressful and even traumatic for many children, especially those who cannot read or understand English proficiently. Students who are spending much of their time trying to read and comprehend questions and answer choices can experience significant anxiety as time passes and they have completed relatively few items. We hypothesize that providing students the opportunity to listen to questions and answer choices in English, and the option to listen to questions and answer choices in their first language, will engage students in less stress and more enjoyable test-taking experiences, and produce more accurate, insightful and higher estimates of science learning and achievement.

***Goals and Aims***

The overarching goal of the proposed research is to investigate our hypotheses that spoken assessments will produce more engaging test-taking experiences, and result in more accurate and higher estimates of students' of students' science knowledge and abilities. The research has two specific aims. The first aim is to test the hypothesis that enabling students to listen to spoken questions and answer choices in English will produce higher estimates of science learning than receiving written choices alone. To test this hypothesis (Aim 1), we will compare performance of students on a standardized computer-administered written assessment that is distributed with the FOSS (Full Option Science System) program, used by over one million children in all 50 U.S. states, to a version of the test that incorporates spoken presentation of the written questions and answer choices. In this comparison, students within classrooms in low- and mid-performing schools who have received identical instruction in an area of science using the inquiry-based FOSS program will be randomly assigned to the written-only or speech-enabled version of the test, in which written questions and answer choices will be highlighted and read aloud to the student. The second hypothesis (Aim 2) is that engaging students in spoken dialogs with a system in which a pedagogical agent is able to conduct in-depth interviews with individual students to assess their understanding of specific areas of science (related to classroom science investigations and instruction), can produce accurate assessments of students' understanding of science and their ability to articulate their understanding through spoken explanations. We will test this hypothesis by comparing spoken dialogs designed to assess students' science understanding administered by a virtual science tutor and by an expert human interviewer.

We believe this research has strong intellectual merit and potentially transformative impact on science education for young learners, and is aligned well with the goals of the PRIME program, which "calls for studies with special emphasis on developing innovative STEM evaluation methodologies and identifying ways to measure or demonstrate the impacts of STEM education programs. Approaches are encouraged that address new ways to conceptualize evaluation, such as a focus on themes of national importance (e.g., teacher education, cyberlearning, innovation) rather than on particular projects or programs." Our research presents a new conceptualization for measuring young learners' science understanding that emphasizes listening comprehension rather than reading comprehension and will include assessment of both conceptual and procedural knowledge.

**Specific Aim 1:** Investigate the benefits of incorporating spoken prompts into written assessments of science knowledge.

Do spoken assessments of science understanding provide more engaging and effective measures of science learning than written ones for most students? We will test this by highlighting all written instructions, questions and response choices in existing written Assessment of Science Knowledge (ASK) items used in the Full Option Science System (FOSS), described below. Fourth grade students will be administered either the existing computer-based written assessment or a new version of the test in which the written instructions, questions and answer choices are highlighted and spoken aloud by a virtual science teacher. Students will be able to control the pace of the assessment, and click on a "repeat" icon to hear the virtual teacher repeat questions or answer choices. In addition, we will investigate the expected benefits of providing first language support to students who speak Spanish (representing over 98% of students with limited English proficiency in our student population); students will have the option to click on icon to listen to a spoken Spanish translation of each English utterance produced by the virtual teacher.

We hypothesize that the large majority of students in low- and mid-performing schools (based on school performance on Colorado state science (CSAP) tests administered in fifth grade) will benefit from spoken assessments, and that the nature of the effect will depend upon individual students' English reading and language skills, which will be measured independently before study begins in a FOSS science module, and will be used as covariates in the analyses. Our hypothesis is that students who receive spoken language support will reveal greater understanding of science on the speech-enabled version of the post-module summative assessment. This hypothesis is based on the view that written English assessments, spoken English assessments, and spoken English assessments with first language support for English learners form a continuum for assessing students' science knowledge with increasing accuracy. For example, an English learner with limited English reading and language proficiency would be expected to demonstrate the least understanding of science using a written English assessment, better understanding when questions and answer choices are presented in spoken English (since they have listened to English during classroom science investigations and instruction), and even better understanding when presented with spoken questions and response choices in their first language (since they are likely to understand the question and answer choices more accurately in their first language). We will investigate the benefits of spoken assessments under these conditions.

**Specific Aim 2:** Can spoken interviews, in which a virtual or human teacher elicits explanations from, and engages a student in, a dialog designed to probe the student's ability to explain science, produce accurate assessments of students' science understanding? In Aim 2, we will test the hypothesis that spoken dialogs between a student and a virtual or human teacher, in which the student is asked to explain the science presented via media (illustrations, animations, or interactive media), can produce detailed and accurate assessments of individual student's science understanding. We hypothesize, based on the research discussed below, that spoken dialogs between a virtual or human science teacher and a student can be designed to elicit spoken explanations from the student that, over the course of the dialog, provide a detailed and accurate portrait of the student's understanding of the science being discussed. That is, the spoken responses during the dialog will reveal whether the student is able to construct accurate explanations, and if not, reveal what the student does and does not understand about the science. We will argue that our prior research demonstrates that speech recognition, natural language understanding and dialog modeling technologies have matured to the point where spoken dialogs can be used efficiently and effectively to assess students' conceptual and procedural knowledge, and that such dialogs produce engaging and exciting experiences for the vast majority of students.

### *The Importance of Scientific Discourse in Science Instruction and Assessment*

The spoken assessments we propose to develop and evaluate are designed to provide teachers and students with powerful tools that can be integrated into science curricula to meet the learning goals described in the New Science Framework [5], which provides the basis for the National Science Standards scheduled for release in 2012. The New Science Framework is consistent with the conclusions and recommendations of the National Research Council Report "Taking Science to School: Learning and Teaching Science in Grades K-8"[6].

The NRC report highlights crucial principles of scientific proficiency: "Students who are proficient in science: 1. know, use, and interpret scientific explanations of the natural world; 2. generate and evaluate scientific evidence and explanations; 3. understand the nature and development of scientific knowledge; and 4. participate productively in scientific practices and discourse" [6, p.2]. The report also emphasizes that scientific inquiry and discourse is a learned skill: "The norms of scientific argument, explanation, and the evaluation of evidence differ from those in everyday life. Students need support to learn appropriate norms and language for productive participation in the discourses of science." [6, p.2].

These principles speak to the role of discourse in building scientific knowledge. Increasingly sophisticated discourse, in which students acquire a deep understanding of science by sharing, reflecting on and refining their ideas as they consider questions, and answer choices and spoken answers, is a primary focus of this proposal. The *2011 Science Framework* incorporates these components of scientific proficiency as it "focuses on important practices used by scientists and engineers, such as modeling, developing explanations or solutions, and engaging in argumentation." [5, p.3]. A key theme of our research is that it is efficient, effective and *necessary* to present students with spoken assessments of science knowledge. Spoken assessments are an essential component of a paradigm shift that is required for educators and students to achieve one of the primary goals in the New Science Framework-- to help teachers learn how to facilitate conversations where students learn, exercise and become proficient in scientific discourse and argumentation. This can only happen if assessments of students' science knowledge assess these abilities. While written tests can be designed to provide excellent assessments of students understanding of science concepts, spoken (and written) assessments that are interactive, that process students' responses and provide follow-on questions to probe students' understanding and their ability to explain science, are an essential component of the New Science Framework.

Without question, this requires a significant change versus most current classroom practices. Despite compelling evidence that engaging students in meaningful conversations improves learning [7; 8; 9; 10; 11; 12; 13; 14; 15], teachers in most classrooms rarely give students the chance to express and share their ideas [16; 17; 18] . Classroom observation studies indicate that only 4% of the questions teachers ask fall into the deep-reasoning category of Bloom's taxonomy [19; 20; 21; 22]. As Osborne [23] notes, "Argument and debate are common in science, yet they are virtually absent in science education."

We hypothesize that spoken assessments based on proven principles for effective conversational interaction will be highly effective in engaging students and challenging them to think and reason about science *if they are informed by the theory and empirical research on effective learning through conversational interaction.* Below, we review some of the relevant research in tutoring and approaches to classroom conversations that have been proven effective in engaging students and improving learning.

Under Aim 2, we will investigate hypothesized benefits of incorporating *spoken explanations* into science assessments, as well as the ability of computer speech recognition and natural language understanding systems to measure students' science understanding relative to human scoring. During spoken explanations, the virtual teacher will present a deep reasoning question, which may be accompanied by an illustration or animation, such as, "Tell me what will happen when the switch is closed in this electric circuit." Based on analysis of the student's explanation, follow-on questions will be generated. We believe it is extremely important to incorporate students' spoken explanations into science assessments to achieve the goals of the New Science Framework, for several reasons. First, there is a strong emphasis in the new Science Framework on having students learn to talk about and explain science. It is therefore important to evaluate their proficiency in this area in standardized assessments. Incorporating assessment of spoken explanations into high-stakes tests is likely to motivate teachers to incorporate science discussions into their classroom activities because teachers are increasingly accountable for their students' performance and have learned to align classroom instruction with state standards and tests. Third, we expect that analysis of spoken explanations during spoken assessments with Marni will be shown to provide greater insights about students' science understanding, and provide new insights on how to design measures for evaluating student's ability to construct and articulate scientific explanations.

## 2. Educational Context

Three key elements of the project's context include Boulder Valley School District (BVSD), its Summer Science Camp (SSC), and the Full Option Science System (FOSS).

Boulder Valley School District: BVSD has 27,000 students, 34 elementary schools, 17 middle schools, and 11 high schools. Selected classrooms in low-and mid-performing schools will be used to conduct field tests and the proposed feasibility study. Our research team and BVSD have worked closely together in previous research. Classrooms in BVSD have high-speed Internet connectivity and the teachers and students participating in the field tests and evaluation studies will be outfitted with appropriate equipment. All but two BVSD elementary schools use FOSS. We will work closely with BVSD administrators and teachers during development, field testing, and assessment of spoken assessments.

BVSD Summer Science Camp (SSC): During the past three years, BVSD has run a 5 week summer program for language-minority students, with over 1000 students attending each summer [24]. Students were invited to attend the 2010 Summer Science Camp if they were non-English proficient (NEP) or limited English proficient (LEP) on the Colorado English Language Assessment, or if they qualified for Extended School Year (ESY) services. The camp provides an ideal test bed for introducing and testing technology-based innovations with English learners and students with special needs and a unique educational context for refining Spanish language support in CASUM dialogs and spoken assessments of science knowledge with Marni.

FOSS: is an inquiry-based science program in use in every state in the United States by over 100,000 teachers and 2 million students. It is used in half of the 100 largest U.S. school districts. FOSS has been developed since 1988 at Berkeley's Lawrence Hall of Science, with support from multiple NSF grants. Twenty-six grade K-6 modules have been developed.

FOSS is a strong match for our research because our team has acquired a deep knowledge of the program during the MyST project (described below), and have a strong relationship with BVSD. The program is widely used, and is distributed with computer-administered summative assessments with strong arguments for validity and reliability. We emphasize that while spoken assessments and dialogs will be developed and tested in the context of FOSS for the purposes of this research, the research results and work products resulting from the proposed study will be designed and intended for use with any science curriculum, as the spoken questions, answer choices and dialogs are aligned with state and national science standards in the New Science Framework, and do not incorporate scenarios or digital content that is FOSS-Centric.

## 3. Theoretical and Empirical Rationale

The project integrates strong theoretical and empirical frameworks. It explores a novel blend of technologies and approaches that combines and responds to multiple R&D strands that have proven effective in separate contexts. This section reviews three of these strands, including social constructivism, classroom dialog and self-explanation, and classroom media use.

What are the factors that make tutoring and classroom dialogs so effective:

1. Question generation: A significant body of research shows that learning improves when teachers and students ask authentic, deep-level-reasoning questions [20]. Asking authentic questions leads to improved comprehension, learning, and retention of texts and lectures by college students [25; 26; 27] and school children [11; 13; 28].

2. Self-explanation: Research has demonstrated that having students produce explanations improves learning [11; 13; 28; 29]. Hausmann and Van Lehn [30; 31] note that: "self-explaining has consistently been shown to be effective in producing robust learning gains in the laboratory and in the classroom."

3. Knowledge co-construction: Students co-construct knowledge when they are provided the opportunity to express their ideas, and to evaluate their thoughts in terms of ideas presented by others. There is compelling evidence that engaging students in meaningful conversations improves learning [7; 8; 11; 13; 14; 15; 28; 29]. Tutorial dialogs increase the opportunity for occurrences of knowledge co-construction, which has been shown to have a significant impact on learning gains [28; 32; 33; 34].

Benefits of multimedia presentations: The integration of narrated animations and interactive media into spoken dialogs (Aim 2) is informed by research on multimedia learning conducted by Richard Mayer and his colleagues [See 35 ; 36, for reviews]. This work investigated students' ability to learn how things work when information was presented in different modalities; e.g., text only, narration of the text only, text with illustrations, narrations with sequences of illustrations, or narrated animations. A key finding of

Mayer's work is that simultaneously presenting spoken narration with visual information results in the highest retention of information and application of knowledge to new tasks. Mayer argues that in a narrated animation, a student's auditory and visual modalities are processed independently but are integrated to produce an enriched mental representation. In spoken assessments, we will integrate illustrations, animations and interactive assessments as a way of focusing the student's attention on the science to be explained, and to help them visualize the science they are talking about. In effect, the open-ended questions provided by the virtual teacher are designed to scaffold thinking and reasoning and help the student construct and articulate an explanation of the science they are viewing. Below, we discuss the dialog principles that are used to facilitate this process.

## 4. Results of Prior NSF Support

My Science Tutor (MyST) is an intelligent tutoring system was developed under NSF DRL 0733323, 9/20/2007 – 8/31/2012 (Wayne Ward – PI, Ron Cole – Co-PI) to improve science learning by third, fourth and fifth grade students through spoken dialogs with Marni, a virtual science tutor, in multimedia environments. Details of the MyST program are described in Ward et al. [37; 38]. Dialogs with Marni continuously assess a student's state of knowledge based on their spoken responses, and scaffold learning by providing students with sufficient information (through questions and media) to challenge them to acquire new knowledge [39; 40]. Each tutorial dialog is oriented around key concepts the student is expected to learn from classroom instructional activities related to FOSS science investigations. The goal of MyST dialogs is to help students construct and generate explanations that express their ideas.

How MyST Works: MyST utilizes automatic speech recognition, character animation, robust semantic parsing, dialog modeling and language and speech generation to support conversations with Marni. The key points of a dialog are specified as propositions realized as semantic frames. During spoken dialogs, the tutor presents illustrations and animations, and asks open-ended questions to elicit speech that entails the targeted propositions. When students respond to Marni's questions, the speech recognition system produces word strings that are processed by semantic grammars to fill the semantic frames that represent the concepts and the relationships among concepts that the student has expressed. Throughout a dialog, the system analyzes utterances produced by the student and maintains a context that represents which points have been addressed by the student, which have not, which were expressed correctly and which represented misconceptions. Based on the current context, the system generates questions to elicit explanations that incorporate the concepts needed to produce a complete explanation. Follow-up questions and media presentations are designed to scaffold learning by providing hints about the important elements of the investigation that the student did not include. The follow-up questions are created by taking a relevant part of the student's response and asking for elaboration, explanation, or connections to other ideas. Details of how MyST works are provided in Ward et al. [37; 38].

MyST Development: Spoken dialogs in MyST were modeled on dialogs between expert human tutors and students. The MyST project tutors were trained in a proven approach to classroom dialogs called Questioning the Author, described below. The first stage of the development process involved analyzing videos of human tutors and students discussing science experiences and phenomena encountered in classroom science investigations in the FOSS program. Reviews of these dialogs were used to develop illustrations, animations and interactive simulations that were tested and refined in subsequent dialogs with human tutors. These sessions were analyzed, transcribed, and used to refine the media, which were then implemented within the MyST system. Dialogs were then refined through a series of field tests with up to 100 students in each test, during which students interacted with the virtual tutor Marni in Wizard of Oz (WOZ) mode. In this mode, Marni and individual students engaged in spoken dialogs, while a human Wizard (a project tutor) in a remote location (our lab at BLT) monitored the session. The Wizard was able to view the student's screen, listen to Marni and the student speak, and view the next response the system was about to produce—the question Marni was about to produce, and the media that would be presented. The teacher had the option of accepting the response, or typing in a new response and presenting a different graphic. These sessions were transcribed and analyzed over the course of the field tests, and used to improve the accuracy and coverage of the speech recognition and natural language understanding systems, and to refine dialog prompts. This same approach will be used to develop spoken dialogs for assessing science understanding (Aim 2), as described in the Development section below.

*How Marni facilitates Conversations using Questioning the Author (QtA).* Conversations with Marni use a proven approach to productive discourse in classrooms called Questioning the Author (QtA), developed by Beck and McKeown [41; 42; 43; 44]. QtA is a scientifically-based and effective program used by hundreds of teachers in the U.S. Recent studies [45; 46] identified QtA as one of two approaches out of nine examined that is likely to promote high-level thinking and comprehension of text. Relative to control conditions, QtA showed effect sizes of 0.63 on text comprehension measures, and of 2.5 on researcher-developed measures of critical thinking/reasoning [45]. Moreover, analysis of QtA discourse showed a relatively high incidence all indicators of productive discussions likely to promote learning [8; 17; 47].

QtA's focus is to have students grapple with, and reflect on, what an author is trying to say to build a representation from it. In the context of an inquiry-based science program, the perspective of the "author" in "Questioning the Author" moves from questions about what a specific author is trying to communicate to questions about science investigations, science phenomena and outcomes. In a sense, the "author" is Mother Nature, and the "texts" are the observations and data sets and ideas that accrue from science investigations and classroom activities. During the past 5 years, in the context of the My Science Tutor project described below, our research team worked with QtA co-developer Margaret McKeown, to create a set of sixteen 20-minute tutorial dialogs in four areas of science in which Marni incorporated media and QtA-based dialog moves into natural spoken dialogs with children to improve their understanding of science. The successful outcomes of this research are described below.

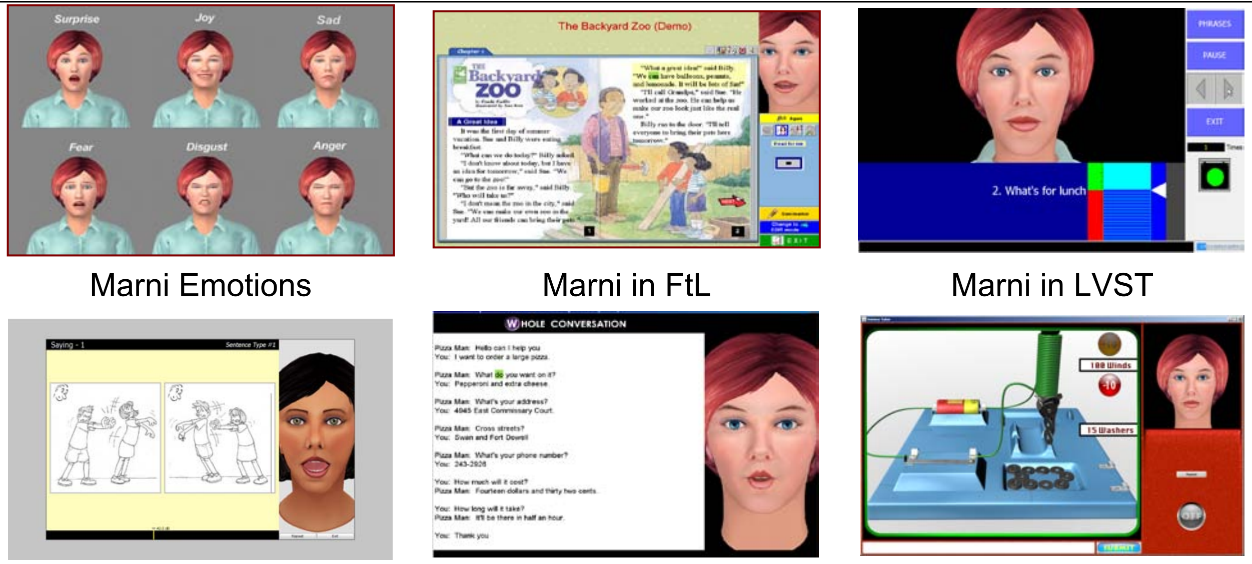
MyST Evaluation: During the 2010-2011 school year, the MyST program was evaluated by comparing learning gains of students who received tutoring sessions with either the virtual tutor Marni (MyST) or with human tutors in small groups. Students were randomly assigned within classrooms to the tutoring condition (virtual or human), and these groups were compared with students from intact control classrooms using FOSS. Eighty-three students received MyST tutoring, 69 were human tutored, and 1015 students in 50 classrooms received classroom instruction and no supplemental tutoring. The FOSS summative ASK assessments were used to measure learning gains for each of the four modules in the study. The hypotheses for the study were: 1) students in MyST and human-tutored groups would have roughly similar gains from pre to post test, 2) tutored students would have significantly greater gains than students in the control (non-treatment) condition.

Details of the evaluation are provided in Ward et al [38]. The results supported the hypotheses: Significant learning gains were found between students tutored by Marni and the control group (d =.53), and human-tutored students and the control group (d = .68). Direct comparisons for the tutoring groups (MyST and Human Tutored) showed no significant differences between groups with t = -1.14, df = 150, p = 0.25. Analysis of surveys completed by students and their teachers indicated that students had highly positive experiences in both tutoring conditions, with students reporting that they felt Marni helped them learn science and were more excited about science because of the tutoring. Teachers also had highly positive impressions of the system, reporting that it had positive impact on their students, that they would use MyST in the future if it were available, and they would recommend it to other teachers. Histograms of students' and teachers' responses to survey questions can be viewed at [48].

***Marni: Fostering Student Engagement, Motivation, Trust, Caring and Learning***

Marni is a lifelike computer character who has served as the face, voice and personality in five learning programs. During the past ten years, Marni has been a reading tutor, a science tutor and a speech and language clinician in projects funded by the NSF, NIH and IES [49; 50; 51; 52; 53]. Our overarching goal when developing each of these programs was to produce qualitative experiences and learning outcomes comparable to those achieved by an expert human tutor or clinician by modeling their speech and language behaviors during interactions with students or clients. In each program, Marni's voice was recorded by an experienced tutor or clinician, so Marni, to the extent possible, assumed the personality and emotions communicated by the person who made the recordings.

Marni is controlled by the CU Animate system, developed by Dr. Jiyong Ma under an NSF-ITR grant (Ron Cole, PI) [54; 55; 56; 57]. Given an input text string and a digitized recording of the words in the text string, CU animated generates accurate visual speech produced by Marni that is synchronized with the recorded utterance. This is accomplished using a multi-unit search algorithm that locates and concatenates the largest sequences of motion capture data in a corpus collected from a speaker with markers placed on her lips and face who produced words, phrases and sentences containing all sequences of English phonemes. The motion capture data is then processed to align the motion capture data to the phonemes in the recorded utterance, and the motion marker positions from the speaker's lips are used to move the vertices of the polygons on Marni's face/ in synchrony with the recorded speech. The result is natural and anatomically accurate visual speech synchronized with each word in a recorded utterance. Marni produces natural movements of the head and face (blinks, eyebrow raises) while speaking, listening to a user speak, or waiting for a user to respond. Six basic emotions can be invoked before, during or after an utterance. Marni currently speaks English, Chinese, Polish and Spanish in existing programs, and "understands" English and Spanish using the speech and natural language understanding processing systems described in the Prior Work section.



Pizza Man: Hello can I help you
You: I want to order a large pizza.

Pizza Man: What do you want on it?
You: Pepperoni and extra cheese.

Pizza Man: What's your address?
You: 4045 East Commissary Court.

Pizza Man: Cross streets?
You: Swan and Fort Dowell

Pizza Man: What's your phone number?
You: 243-2926

You: How much will it cost?
Pizza Man: Fourteen dollars and thirty two cents.

You: How long will it take?
Pizza Man: It'll be there in half an hour.

You: Thank you

Marni Emotions

Marni in FtL

Marni in LVST

Sentactics

Marni in Aphasia Scripts

Marni in MyST

In the Foundations to Literacy (FtL) program (2003-2006), Marni served as a reading tutor who interacted with over 2000 kindergarten, 1st, 2nd, and 3rd grade students in Colorado inner city, suburban and rural schools to help them acquire foundational reading skills and exercise these skills in interactive books to become fluent readers. Students improved letter recognition, word reading and passage comprehension skills on standardized tests [49]. During the past five years, in the context of the My Science Tutor program, Marni has tutored over 500 3rd, 4th and 5th grade students in both WOZ studies and during independent use in summative evaluations. In both studies, students were administered surveys; K-2 students using FtL were read questions and response choices out loud, and students using MyST were presented written surveys. Classroom teachers were also administered written surveys. Responses to most questions used a three point Likert scale, with questions like: "Was Marni a good reading teacher?" with responses like "Marni was a very good reading teacher," Did help you learn how to read? And choices like "Marni was a great reading teacher," Marni was sometimes a great reading teacher," "Marni was not a good reading teacher." Both student and teacher responses were highly positive in both FtL and MyST with over 95% of all responses in the top or middle response categories, and approximately 50% of all responses receiving the most positive rating across all questions. Perhaps most interested, it was clear that students bonded with Marni and interacted with her as if she was a real person. For example, students reported that they trusted Marni, and that the believed Marni cared about them. Teachers also produced highly positive ratings of both programs, reporting that they helped children learn, that they would use the program again, and would recommend it to other teachers. Our results are consistent with studies by Reeves and Nass [58] summarized in *The Media Equation*, which demonstrate that people interact with computers using the same rules and social conventions they use when interacting with other people, and with studies by Lester [59], who coined the term "Persona Effect" to describe the observation that students form a bond with pedagogical agents in well-designed learning programs, and are motivated to work hard to please the agent. Histograms of teachers' and students' survey responses after using FtL and MyST can be viewed at [48]. We view the positive experiences that teachers, children and adults have had with Marni as evidence that children are likely to become engaged with Marni during spoken assessments, resulting in less stressful and more enjoyable test-taking experiences. We plan to document this hypothesis by designing surveys similar to those in our previous studies, and by observing students using the system.

## 5. Research, Development & Assessment Plan

### *Development Plan*

The Speech-enabled computer-administered FOSS ASK assessments (Aim 1) and spoken dialogs for assessing conceptual and procedural science knowledge (Aim 2) will be developed for one fourth grade FOSS module. The specific FOSS module will be selected by Dr. Samantha Messier, Director of Science Curricula at BVSD, from the new 2012 FOSS release which is aligned to state and national standards in the New Science Framework. (Please see letters of support).

All development activities will use participatory design methodology, in which teachers collaborate as integral part of the software development process from the inception of the project. Two teachers will be recruited to work as part of the project team. They will provide input and feedback on the initial design and development of the spoken assessments, observe students using the assessments initially in the laboratory (at BLT), and then in field tests in their classrooms following instruction in the targeted module. During field tests, involving two to three classrooms (but excluding classrooms of teachers in the participatory design team), we will observe students while taking the spoken assessments, observe and interview teachers and students, and analyze system logs that record all student and system responses. The field tests will be used to refine the user interface as needed for speech-enabled FOSS ASK assessments, and to refine the task grammars, semantic frames and dialog moves made by the virtual teacher during spoken dialogs. Although FOSS modules have been revised for the new 2012 FOSS release, which will be used in our study, much of the content in the modules is the same as the previous FOSS release; we plan to leverage the media and dialogs developed during the MyST project. We note that our research team has deep knowledge of the FOSS program and the 2012 release; BLT has worked with the FOSS developers at Lawrence Hall of Science to develop interactive tutorials and virtual investigations that will be distributed for each of the FOSS modules for grades 3-6 in the 2012 FOSS release.

**Speech-enabled ASK assessments:** Incorporating spoken instructions, questions, answer choices and other prompts into the written computer-based FOSS assessments is relatively straightforward, as all of the infrastructure has been developed and tested as part of the MyST project. In the current evaluation of MyST (now underway), students conclude each dialog with Marni with a deep reasoning question that is presented as text, highlighted, and spoken by Marni. The question is accompanied by an illustration or animation. After students produce a spoken explanation to the question (which is recorded), they are presented with printed response choices that are highlighted and read out loud. This infrastructure will be modified and refined for integration into the written ASK assessments. As noted above, having Marni produce speech in either English or Spanish is also straightforward. A bilingual voice talent will be auditioned, and will practice reading the prompts out loud. The prompts will then be recorded in English and in Spanish. The CU Animate system automatically generates Marni's visual speech and head movements, and synchronizes the visual speech with the recordings. The project staff, and the BVSD translation services department, will review and approve all translations. Prompts may be re-recorded based on the results of laboratory user testing and the field tests. Students will be able to repeat prompts in English and Spanish during field testing and summative evaluation. Prompts that are repeated more often than others by students during field tests will be reviewed and re-recorded as necessary.

Spanish language support for ASK assessments: Our studies with MyST indicate that English learners students become fully engaged in dialogs with Marni, and that learning gains in schools with the most English learners produced the highest learning gains in the summative evaluation of MyST, which is highly encouraging. Our plan is to provide all students the option of clicking on an icon to listen to a Spanish translation of the last utterance that Marni produced in English. This will enable us to investigate how often students select this option, and correlate the frequency of selection of Spanish prompts with independent assessments of reading and English language proficiency administered before students begin a FOSS module, and to investigate how frequency of selection of Spanish prompts correlates with learning gains. We plan to use the same procedure for Spanish language support during Spoken dialogs with Marni to assess science understanding. Field testing of both spoken ASK assessments and spoken dialogs with Spanish language support will be conducted in the BVSD Summer Science Camp with English learners and special needs students in third, fourth and fifth grades, and during field testing in participating classrooms during the regular academic year.

Spoken Dialogs for Assessing Science Understanding: Based on the infrastructure, experience and reusable content developed during the MyST project, we plan to develop and test spoken dialogs within the MyST system to assess science learning in the selected FOSS science module. The dialogs will be tested initially with students in the laboratory who have recently completed the FOSS modules during the first year of the project. As soon as user testing is completed, field tests begin. We note that FOSS modules are taught during each quarter of the academic year, enabling field tests to be conducted throughout the year.

During the first field test, we will use the WOZ paradigm, described above, to collect data that will be transcribed and used to refine the spoken dialogs. We estimate that it will take approximately two months to develop spoken dialogs, enabling two rounds of field tests, with up to three classrooms in each test, before spoken dialogs are evaluated during the summative evaluation in the Spring and Winter quarters of year two.

*Evaluation Plan*

Structure of Evaluation Section: In this section, we will describe our two studies. Then, we will describe each of the design components that will be common to both studies to avoid repeating the same details. Lastly, we will present analytic plans that are tailored to each study. Both studies will be conducted in the last semester of the project when students use the module for which we will have developed the spoken assessments and dialogues.

*Overall Aims*: The research has two specific aims. The first aim is to test the hypothesis that enabling students to listen to spoken questions and answer choices in English, with the additional option of listening to spoken presentation of questions and answer choices in Spanish, will produce more accurate estimates than of science learning than receiving written choices alone. To test this hypothesis (Aim 1), we will compare the performance of students on a standardized computer-administered written assessment associated with the FOSS program (ASK assessments, described in much greater detail below). In this comparison, students within classrooms in low- and mid-performing schools who have received identical instruction in an area of science using the inquiry-based FOSS program will be randomly assigned to the written-only or speech-enabled version of the test, in which written questions and answer choices will be highlighted and read aloud to the student.

The second hypothesis is that engaging students in spoken dialogs with a system in which a virtual, pedagogical agent is able to conduct an in-depth interviews with individual students to assess their understanding of specific areas of science (related to classroom science investigations and instruction), produces accurate assessments of student understanding of science and their ability to articulate it through spoken explanations. We will test this hypothesis (Aim 2), by comparing spoken dialogs designed to assess students' science understanding administered by a virtual science tutor, also provided by an expert human interviewer. We will seek to find a difference in how the interviews are conducted, and if the virtual interview can approach the accuracy of measurement of expert human interviewers.

Participants' selection and assignment: Throughout the two studies we will randomly select students from the pool of willing participants and randomly assign those students to treatment or nontreatment (control) conditions. We do, however, have a problem with random assignment with so few students in each classroom. In essence, our attempt to randomly assign students might not result in randomly constructed groups. To examine this, with STATA – PSMATCH2, we will use propensity score matching on pretests (assessments of science, reading, and language that each student will be given as part of their standard curriculum) and the Mahalanobis Metric Matching method [60], comparing the propensity score constructed groups to the randomly selected groups. This will help us determine if we need to block students on pretest or language test scores before randomly assigning them to experimental groups. If we are unable to block students, this analysis will at least provide us with important covariates for later analyses (described below).

Student and school characteristics: In conjunction with the methods of qualitative data analysis, we will also collect available information on schools, ranging from broad demographic information to school contextual measures [61].These data will help us understand the context from which treatment and non-treatment students come. Analytically, it will help control for school- and teacher-level differences to minimize bias in the treatment and non-treatment group comparisons.

We will also collect demographic information on the students themselves. For example, we will collect information on students' free or reduced lunch rate status, sex, ethnicity, disability classification, etc. We will use the information to control for differences between treatment and non-treatment students. Accordingly, we will use this information to generate potential covariates that may or may not vary over time.

Measures: When investigating student science learning in the first study we will use the ASK assessments varying their modality. These assessments are associated with the FOSS curriculum. The ASK (Assessing Science Knowledge) consists of a set of summative assessments administered to students before and after each science module. In addition, students' entries in science notebooks, which take place during each classroom science investigation, can be scored for science knowledge using established rubrics. The pre and post-modules summative assessments have between 8 and 12 items and show composite internal reliability with alphas in the 0.80's and 0.90's range. The interrater reliability for subjective items has also met high standards in similar conditions (e.g., r = 0.90), and the validity of the measures has been built up over time through a process of empirical investigation.

Second, we will use BVSD district test scores of reading (administered two to three times per year to all at risk students in 3rd-5th grades by teachers, and to all participating students by project staff in the pilot studies). Teachers who are trained to administer this test achieve an inter-rater reliability of 0.80. The DRA2 is a criterion-referenced test of reading fluency and comprehension. The assessment uses *Benchmark Books* (by reading difficulty), which are given to students to read in front of the teacher and alone. Typically, K – 3rd students are asked to retell the story, summarizing what they read, while 4th - 8th grade students respond to the assigned texts in writing. As part of our study, we will further require that all students verbally retell the stories in addition to their written prompts. We will gather raw student scores to determine pretest covariates of language. Teachers use rubrics to score student responses on dimensions of oral reading, fluency, and comprehension.

In addition, tests of oral reading fluency will be administered using EasyCBM [62], a Curriculum Based Measurement and assessment system currently used by over 170,000 teachers across 50 states, with well over 7.7 million reading and math tests taken by students in grades K-8. Developed in 2006, EasyCBM is an online benchmarking and progress monitoring assessment system available for free teacher and researcher use. EasyCBM includes multiple different reading forms for students in grades K-8, with appropriate difficulty within each grade. Measures of oral reading fluency use text passages that are consistent in their length and readability at each grade-level using the Flesch-Kincaid index feature available on Microsoft Word [63]. Tests from EasyCBM will be used to measure word reading and word frequency for lower grades, and word reading, frequency and vocabulary (both proximal with instruction and distal from instruction) for grades three to five. Computer administered EasyCBM tests have established validity and reliability arguments and offer nine equivalent, but non-identical forms for pre/post comparisons. Word reading, frequency and vocabulary tests are short (three minutes each) and adapt to the initial reading level of the student. The Gates-Macginitie (a group administered paper and pencil test) will assess reading comprehension.

Oral reading fluency will be measured automatically using FLORA, a system developed at BLT, which will be integrated into EasyCBM, which administers fully automated tests of oral reading fluency [64; 65]. Oral reading fluency has been demonstrated to be a valid and reliable measure of reading proficiency that correlates highly with comprehension of texts and students' future reading progress. Research has demonstrated that FLORA produces two accurate measures or oral reading fluency-- *Words Correct Per Minute (WCPM),* a standardized measure of oral reading fluency with published norms for each grade level[66], and a measure of *how expressively* the student reads a text passage, using the standardized 4 point NAEP scale [3]. Research by Bolenos, et al [64; 65]demonstrated that scores produced by FLORA have the same level of agreement on a corpus of over 700 stories read by over 300 children as the inter-rater agreement between trained human judges. We will use FLORA to measure oral reading fluency for every student. The administration will provide for us covariates associated with language proficiency.

Finally, all native Spanish-speaking students who participate in the study (the vast majority of English language learners (ELLs) in Boulder Valley School District) will be administered the Colorado English Language Acquisition (CELA) pre/post, which will be used to model any changes to language learning occurring with ELLs over the course of the intervention.

Analytic Plans: For each of the two research hypotheses (with rationales provided above), we will perform analyses using the same Structural Equation Modeling (SEM) techniques. Discussions of the analyses for each of the hypotheses are described below.

*Hypothesis 1: Students who listen to spoken questions and answer choices in English, with the additional option of listening to them in Spanish, will produce more accurate and comprehensive estimates of science knowledge than students that receive written questions alone.*

To test this hypothesis we will compare performance of students who undergo a standardized, computer-administered written assessment with students who undergo a computer-administered spoken assessment. In this comparison, students within classrooms in low- and mid-performing schools who have received identical instruction in an area of science using the inquiry-based FOSS program will be randomly assigned to the written-only or speech-enabled version of the test, in which written questions and answer choices will be highlighted and read aloud to the student. As already stated, we will ensure groups are equivalent using blocking variables, or even propensity score analyses.

To evaluate our results we will use a structural equations modeling (SEM) approach. Following the work of Embretson [67] and Kline [68], we will use their seven steps of model building. We chose SEM for several reasons, but the primary reason was that we are extremely unlikely to meet the restrictive assumptions of standard linear modeling techniques (e.g., ANCOVA). SEM allows us to throw off these assumptions and address the data as they are, creating a more accurate model of student knowledge, and the differences produced by varying the modalities of our assessments.

Primarily, we are evaluating the effect of written verses written plus spoken language in the assessments. This serves three purposes. First, this evaluation (with the language test covariates) will isolate the effect of hearing and seeing versus only seeing written, multiple choice questions. Second, if our expectations are met, that the oral assessments produce more accurate measures of student knowledge, we will use these oral assessments to strength the ASK assessments themselves and improve the substantive and structural validity arguments associated with them. Similarly, we will improve the ASK assessments as well as their reliability arguments insofar as the student response patterns are more accurately modeled by our IRT models [67].

We will collect data from 96 students (approximately 5 classrooms of students). This was determined by a power analysis using Optimal Design Software assuming: 1) an acceptable alpha of 0.05; 2) a moderate to large effect size (0.60); and 3) a moderate amount of variance explained by covariates. It also assumes a larger than the expected 10% attrition rate to ensure that comparisons are adequately powered (i.e., having a power of 0.80).

*Hypothesis 2: Students that engage in spoken dialogs with automatic interviews will more accurately and comprehensively report their understanding of science (related to classroom science investigations and instruction), resulting in generally higher estimates of student learning.*

We will test this hypothesis by comparing spoken dialogs designed to assess students' science understanding administered by a virtual science tutor and by an expert human interviewer. The analytic plan associated with this hypothesis will be quite straightforward. We will ensure comparable groups using the assignment procedures (described above), confirmed with propensity score analyses.

Again, we will use a structural equations modeling (SEM) approach. Following the work of Embretson [67] and Kline [68], we will use their seven steps of model building. We chose SEM for several reasons, but the primary reason was that we are extremely unlikely to meet the restrictive assumptions of standard linear modeling techniques (e.g., ANCOVA), as already noted. Second, in this study, due to classroom–level variability in how FOSS is presented to students, or natural variation in teacher instruction/disposition, we may have to use classroom-level covariates. SEM not only allows avoiding the restrictive assumptions of classical evaluation techniques, but also building into our model measurement error and multilevel covariates.

We will collect data from 96 students (approximately 5 classrooms of students). This was determined by a power analysis using Optimal Design Software assuming: 1) an acceptable alpha of 0.05; 2) a moderate to large effect size (0.60); and 3) a moderate amount of variance explained by covariates. It also assumes a larger than the expected 10% attrition rate to ensure that comparisons are adequately powered (i.e., having a power of 0.80).

## 6. Management, Timelines, Oversight, and Dissemination

Management: The project will be managed by PI Wayne Ward in close collaboration with co-PI Finbarr Sloane and Brandon Helding. Jeannine Myatt and Cindy Buchenroth-Martin, project leaders in the MyST project, will liaison with BVSD schools, aided by Dr. Samantha Messier, Director of Science Curricula at BVSD. This team worked together to develop and evaluate the MyST system.

Timelines & Oversight: Table1 presents timelines for the main tasks. Oversight and guidance for the project will be provided by an advisory board composed of distinguished educational researchers Kathy Long and Samantha Messier.

Dissemination: The results of this project will be communicated to STEM professionals and practitioners via a project Web site that will be set up and maintained by BLT during the first year of the project. We will present results of the software architecture and applications, field-testing outcomes and assessment plan and outcomes at national and international conferences attended by the computer scientists, engineers, cognitive scientists, and educational researchers who comprise the project team. The results of the project will be described in peer review journals. The assessments instruments will be distributed to the academic community free of charge for research use, and donated to the FOSS team at UC Berkeley for their use.

Expertise: Principal Investigator, Wayne Ward is co-founder and Chief Scientist at Boulder Language Technologies. He will be responsible for managing all aspects of the CASUM project. He co-founded the Center for Spoken Language Research at CU Boulder, where he managed several successful large projects with Ron Cole funded by NSF and DARPA. Wayne is the PI of the MyST project, and developed the MyST system. Co-PI Dr. Finbarr Sloane, a distinguished education researcher and former NSF program director, will manage and conduct the formative and summative evaluation in collaboration with Dr. Brandon Helding. Co-PI, Dr. Brandon Helding received his PhD in Curriculum and Instruction from Arizona State University. He will be responsible for conducting research leading to qualitative and quantitative measures of science knowledge. He has worked as a methodologist on four NSF-funded grants and has published one book on measurement development. Consultant Dr. Timothy Weston will provide substantive external expert review and regular critical review on the project's methods and progress, analysis procedures, interpretation of data into findings, and dissemination activities.

Milestones and Timelines

| | year 1 | | | | year 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | sum 1 | 1-Sep | 1-Dec | 1-Mar | sum 2 | 1-Sep | 1-Dec | 1-Mar |
| **Initial Design** | | | | | | | | |
| Speech-Enabled ASK Assess. | 🟩 | 🟩 | 🟩 | | | | | |
| Develop Virtual Interviews | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | | | |
| Laboratory Testing | 🟩 | 🟩 | 🟩 | | | | | |
| **Field Tests -- Iterative Testing and Refinement** | | | | | | | | |
| Classroom Testing | | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | | |
| Student Interviews | | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | |
| **Pilot Studies** | | | | | | | | |
| Summative Evaluation | | | | | | | 🟦 | 🟦 |

Understanding the Transformative Potential of Spoken Assessments of Science Understanding for Young Learners

## References

1. National Center for Education Statistics. (2011). The Nation's Report Card: Science 2009. Washington, D.C.: Institute of Education Sciences, National Center for Education Statistics.
2. Cromley, J. G. (2009). Reading Achievement and Science Proficiency: International Comparisons from the Programme on International Student Assessment. *Reading Psychology, 30*(2), 89-118.
3. National Center for Education Statistics. (2009). The Nation's Report Card: Reading 2009. Washington, D.C.: Institute of Education Sciences, U.S Department of Education.
4. National Literacy Panel on Language-Minority Children and Youth. (2006). EXECUTIVE SUMMARY: Developing Literacy in Second-Language Learners: Report of the National Literacy Panel on Language-Minority Children and Youth. In D. August & T. Shanahan (Eds.). Mahwah, NJ: Center for Applied Linguistics.
5. Committee on Conceptual Framework for the New K-12 Science Education Standards, & National Research Council. (2011). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*: The National Academies Press.
6. National Research Council. (2007). Taking Science to School: Learning and Teaching Science in Grades K-8. In R. A. Duschl, H. A. Schweingruber & A. W. Shouse (Eds.), *Committee on Science Learning Kindergarten through Eighth Grade*. Washington D.C.: The National Academies Press.
7. Murphy, P., Wilkinson, I., Soter, A., Hennessey, M., & Alexander, J. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101*(3), 740-764.
8. Soter, A., Wilkinson, I., Murphy, P., Rudge, L., Reninger, K., & Edwards, M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research, 47*, 372-391.
9. Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.
10. King, A. (1994). Guiding knowledge construction in the classroom: Effect of teaching children how to question and explain. *American Educational Research Journal, 31*, 338-368.
11. King, A., Staffieri, A., & Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology, 90*(1), 134-152.
12. McNamara, D. S., Levinstein, I. B., & & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavioral Research Methods, Instruments, and Computers*(36), 222-233.
13. Palincsar, A., & Brown, A. (1984). Reciprocal Teaching of Comprehension-Fostering and Comprehension-Monitoring Activities. *Cognition and Instruction, 1*(2), 117-175.
14. Pine, K., & Messer, D. (2000). The effect of explaining another's actions on children's implicit theories of balance. *Cognition and Instruction, 18*(1), 35-52.
15. Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology, 98*(1), 182 -197.

16. Nystrand, M., Gamoran, A., Kachur, R., & Prendergast, C. (1997). *Opening dialogue: understanding the dynamics of language and learning in the English classroom*. New York, NY: Teachers College Press.

17. Nystrand, M., & Gamoran, A. (1991). Instructional Discourse, Student Engagement, and Literature Achievement. *Research in the Teaching of English, 25*(3), 261-290.

18. Cazden, C. (1988). *Classroom discourse: The language of teaching and learning*: Heinemann Portsmouth, NH.

19. Kerry, T. (1987). Classroom questions in England. *Questioning Exchange, 1*, 32-33.

20. Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: the classification of educational goals; Handbook I: Cognitive Domain*. New York, NY: Longmans, Green.

21. Dillon, J. T. (1988). *Questioning and teaching: A manual of practice*. New York, NY: Teachers College Press.

22. Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*(1), 104-137.

23. Osborne, J. (2010). Arguing to Learn in Science: The Role of Collaborative, Critical Discourse. *Science, 328*, 463-466.

24. Messier, S. (2011). Extend, Enrich, and Empower: Closing the Achievement Gap in Science through Summer Learning Opportunities. Retrieved from FOSS, Full Option Science System website: http://lhsfoss.org/newsletters/last/FOSS37.extend.html

25. Craig, S., Gholson, B., Ventura, M., & Graesser, A. (2000). Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education, 11*, 242-225.

26. Driscoll, D., Craig, S., Gholson, B., Ventura, M., Hu, X., & Graesser, A. (2003). Vicarious learning: Effects of overhearing dialog and monologue-like discourse in a virtual tutoring session. *Journal of Educational Computing Research, 29*(4), 431-450.

27. King, A. (1989). Effects of self-questioning training on college students' comprehension of lectures* 1. *Contemporary Educational Psychology, 14*(4), 366-381.

28. King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31*(2), 338.

29. Chi, M., Bassok, M., Lewis, M., Reimann, P., Glaser, R., & Alexander. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145-182.

30. Hausmann, R. G. M., & VanLehn, K. (2007a). Explaining self-explaining: A contrast between content and generation. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), *Artificial Intelligence in Education* (pp. 417-424). Amsterdam, Netherlands: IOS Press.

31. Hausmann, R. G. M., & VanLehn, K. (2007b). *Self-explaining in the classroom: Learning curve evidence.* Paper presented at the 29th Annual Conference of the Cognitive Science Society, Mahwah, NJ.

32. Chapin, S. H., Oconnor, C., & Anderson, N. C. (2003). *Classroom Discussions Using Math Talk to Help Students Learn.Math Solution Publications*. Sausalito, CA: Math Solution Publications.

33. Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*(4), 471-533.

34. Wood, D., & Middleton, D. (1975). A study of assisted problem solving. *British Journal of Psychology, 66*(2), 181-191.

35. Mayer, R. E. (2001). *Multimédia learning*: Cambridge University Press.

36. Mayer, R. (Ed.). (2005). *The Cambridge handbook of multimedia learning*. New York, NY.

37. Ward, W., Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S. V., . . . Becker, L. (2011). My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Trans. Speech Lang. Process., 7*(4). doi: 10.1145/1998384.1998392

38. Ward, W., Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., & Weston, T. (In Press). My Science Tutor: A Conversational Multimedia Virtual Tutor. *Journal of Educational Psychology,* (Special Issue on Advanced Learning Technologies). Retrieved from http://www.bltek.com/images/mindstars/myst_article_submitted_to_jep_special_issue_on_advanced_learning.pdf

39. Vygotsky, L. S. (1978). *Mind in Society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

40. Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163-182.

41. McKeown, M., Beck, I., Hamilton, R., & Kucan, L. (1999). *"Questioning the Author" Accessibles: Easy access resources for classroom challenges*. Bothell, WA: The Wright Group.

42. McKeown, M., & Beck, I. (1999). Getting the discussion started. *Educational Leadership, 57*(3), 25-28.

43. Beck, I. L., & McKeown, M. G. (2006). *Improving Comprehension with Questioning the Author: A Fresh and Expanded View of a Powerful Approach*: Scholastic.

44. Beck, I., & McKeown, M. (2006). Encouraging young children's language interactions with stories. In D. Dickenson & S. Neuman (Eds.), *Handbook of Early Literacy Research* (Vol. 2). New York, NY: Guilford.

45. Murphy, P. K., & Edwards, M. N. (2005). *What the studies tell us: A meta-analysis of discussion approaches, In M. Nystrand (Chair), Making sense of group discussions designed to promote high-level comprehension of texts*. Paper presented at the American Educational Research Association, Montreal, Canada.

46. Murphy, P. K., Wilkinson, I. A. G., Soter, A. O., Hennessey, M. N., & & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' high-level comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101*, 740-764.

47. Soter, A. O., & Rudge, L. (2005). *What the Discourse Tells Us: Talk and Indicators of High-Level Comprehension.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

48. Boulder Langauge Technologies. (2012). BLT-Web, from http://www.bltek.com/mindstars/

49. Cole, R. A., Wise, B., & Van Vuuren, S. (2007). How Marni Teachers Children to Read. *Educational Technology, 24*(1), 14-18.

50. Cole, R., Halpern A., Ramig, L., van Vuuren, S., Ngampatipatpong, N., & Yan, J. (2007). A Virtual Speech Therapist for Individuals with Parkinson Disease. *Educational Technology, 47*(1), 51-55.

51. Cherney, L., Halper, A., Holand, A., & Cole, R. (2008). Computerized Script Training for Aphasia: Preliminary Results. *American Journal of Speech-Language Pathology, 17*, 19-34.

52. Cherney, L. (2010). Oral reading for language in aphasia (ORLA): evaluating the efficacy of computer-delivered therapy in chronic nonfluent aphasia. *Topics in Stroke Rehabilitation, 17*(6), 423-431.

53. Thompson, C. K., Choy, J. J., Holland, A., & Cole, R. (2010). Sentactics®: Computer-Automated Treatment of Underlying Forms. *Aphasiology, 24*(10), 1242-1266.

54. Ma, J., Yan, J., & Cole, R. (2002). *CU Animate tools for enabling conversations with animated characters.* Paper presented at the Seventh International Conference on Spoken Language Processing.

55. Ma, J., Cole, R. A., Pellom, B., Ward, W., & Wise, B. (2004). Accurate Automatic Visible Speech Synthesis of Arbitrary 3D Models Based on Concatenation of Di-Viseme Motion Capture Data. *Journal of Computer Animation and Virtual Worlds, 15*(5), 485-500.

56. Ma, J. a. C., R.A. (2004). Animating Visible Speech and Facial Expressions. *The Visual Computer, 20*, 86-105.

57. Ma, J., Cole, R., Pellom, B., Ward, W., & Wise, B. (2006). Accurate Visible Speech Synthesis Based on Concatenating Variable Length Motion Capture Data. *IEEE Transactions on Visualization and Computer Graphics, 12*(2), 266-276. doi: 10.1109/tvcg.2006.18

58. Reeves, B., & Nass, C. (1996). *The media equation*. New York, NY: Cambridge University Press.

59. Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). *The persona effect: affective impact of animated pedagogical agents*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, Atlanta, Georgia, United States.

60. Rosenbaum, P., & Rubin, D. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association, 79*(387), 516-524. doi: citeulike-article-id:147912

61. Schatzman, L., & Strauss, A. (1973). Field Research: strategies for a natural sociology: Englewood Cliffs, NJ: Prentice-Hall.

62. Oregon, U. o. (2006). *EasyCBM*. Boston, MA: Houghton Mifflin Harcourt.

63. Alonzo, J., & Tindal, G. (2008). The development of fifth-grade passage reading fluency measures in a progress monitoring assessment system. Eugene, OR: University of Oregon.

64. Bolanos, D., Cole, R. A., Ward, W., Borts, E., & Svirsky, E. (2011). FLORA: Fluent oral reading assessment of children's speech. *ACM Trans. Speech Lang. Process., 7*(4), 1-19. doi: 10.1145/1998384.1998390

65. Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., & Schwanenflugel, P. J. (In Press). Human and Automated Assessment of Oral Reading Fluency. *Journal of Educational Psychology*(Special Issue on "Advanced Learning Technologies").

66. Hasbrouck, J., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*, 636-644. doi: 10.1598/RT.59.7.3

67. Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

68. Kline, R. (2005). *Principles and practice of structural equation modeling*: The Guilford Press.