

Proposal Submitted to the National Science Foundation entitled:

Collaborative Research: Improving Science Learning in Inquiry-based Programs

Wayne Ward (Principal Investigator), Ron Cole (Co-Principal Investigator)

Project Summary

The goal of the proposed work is to improve science learning by students who are not achieving their potential in high quality inquiry-based programs. While programs like FOSS, STC and Insights have proven effective in improving science achievement within and across school districts, many children, especially underrepresented minorities and English language learners, fail to demonstrate proficiency on standardized tests of science achievement. The project will aim to achieve its goal by developing a computer program, My Science Tutor, which students will use immediately following classroom science investigations to reinforce and extend concepts embedded in the investigations. The program uses a lifelike animated character to engage students in scaffolded guided learning activities and tutorial dialogs that stimulate and scientific reasoning. Tutorial dialogs are based on a proven technique, Questioning the Author, that challenges students to learn and integrate new concepts with prior knowledge to construct enriched mental models that can be used to explain and predict scientific phenomena. The work aims to produce and demonstrate the effectiveness of tutorial dialogs produced by human experts trained to use the Questioning the Author method. To evaluate the intervention, we will compare learning gains on standardized tests of science achievement by fourth and fifth grade students in four areas of science of student randomly assigned to three groups: the computer treatment, human tutoring or continued classroom instruction. Formative assessments will analyze both student and teacher experiences, and learning of concepts and scientific reasoning through analysis of entries in students' science notebooks.

Intellectual Merit: This study will contribute new knowledge about the influence and impact of well-designed learning tools that are designed to improve concept formation and critical thinking by elementary school students who are not achieving their potential in high quality inquiry-based science programs. The formative assessments should provide detailed insights about how learning tools designed to teach concepts through scaffolded learning and narrated animations, and to teach scientific reasoning through tutorial dialogs, influence the learning and achievement of elementary students. The program will also contribute new knowledge to science about the effectiveness of tutorial dialogs incorporating advanced language technologies to emulate the learning strategies of expert tutors and the learning gains of their students.

Broader Impact: Successful outcomes of the project will produce a program that is effective in improving science learning and achievement of elementary school students. The program will provide an effective supplement to FOSS, a high quality science program that is already used by over two million students and one hundred thousand teachers in the U.S. A potentially profound advantage of the project arises from providing viable and accessible resources to help teachers implement high quality curricula in a much more individualized manner.

INTRODUCTION

The **primary goal** of this project is to develop computer-based learning tools to improve science achievement for low performing students when used as a supplement to high quality inquiry-based science curricula. The program, My Science Tutor, is designed especially for students who do not reach their potential when using these curricula. It will assess students' knowledge of vocabulary and concepts linked to classroom science investigations, and facilitate learning of vocabulary and concepts as needed through contextualized and scaffolded instruction. Additionally, it will facilitate students' thinking and reasoning about these concepts by prompting them to explain and predict scientific phenomena through tutorial dialogs with a virtual tutor. We will evaluate the effectiveness of the learning program on standardized tests of science achievement following random assignment of students to one of three groups for 20 minutes daily following classroom science investigations: one-on-one instruction in My Science Tutor, small group instruction by an expert human tutor, and reading of supplementary materials in a classroom comparison condition. **Successful outcomes** of the proposed work will provide science teachers with resources for rapid and accurate assessment of individual student comprehension of science concepts both to inform instruction and to provide them with effective learning tools that can improve achievement.

The proposed work is closely aligned to Grand Challenges 1 and 3. Our work addresses Grand Challenge 1 by developing "assessments that help teachers diagnose students' comprehension more precisely and accurately and to link good formative assessments to high stakes state tests." Our work addresses Grand Challenge 3 by assembling a "collaborative partnership that involves scientists, mathematicians, engineers, learning scientists, and educators" to infuse new methodologies, technologies and content into K-12 STEM education.

RATIONALE AND OPPORTUNITY

In the 2002 National Assessment of Educational Progress (NAEP), only 2% of U.S. students attained advanced levels of mathematics or science achievement by Grade 12 (EdWeb, 2005). An increasing number of elementary and middle schools in the U.S. are addressing this problem by using high quality inquiry-based science curricula, such as FOSS (Full Option Science System); (FOSS, 2007a), STC (Science and Technology for Children; STC, 2007) and Insights (2007). The FOSS Teacher Guide Introduction explains the program's overall approach: "The best way for students to appreciate the scientific enterprise, learn important scientific concepts, and develop the ability to think well is to actively construct ideas through their own inquiries, investigations and analyses." Programs developed over the past two decades based on this approach are consistent with a large body of theory and research (e.g., Bransford et al., 1999) and with recommendations presented in reports sponsored by the nation's top science and educational organizations (e.g., NRC, 1999; 2001; National Academies, 2006).

While there is clear evidence that student achievement improves in school districts that use high quality inquiry-based science curricula, there are many students, especially those in low performing schools, who fail to achieve their potential. We believe we can improve science education and student

achievement in the United States by developing an intelligent tutoring system addressing specific limitations of inquiry-based science programs that are caused by practical realities encountered in many classrooms. Initially, we will establish a proof-of-concept through FOSS, as one step in developing a sharable methodology for creating similar tutoring systems that can work for other current and yet-to-be developed curricula.

FOSS: FOSS has been under development since 1988 at the Lawrence Hall of Science at the University of California Berkeley with support from three separate NSF grants. Larry Malone and Linda De Lucchi have served as project co-directors and author/developers since the inception of the FOSS Project. Twenty-six modules have been developed for K-6. Each module consists of a kit of student materials, a teacher guide, a module-specific teacher-preparation video, a student reading book, and a website. Within a module, students in classrooms work in small groups to conduct a series of 4 to 6 science investigations over an 8 to 10 week period. Formative assessments are embedded in each investigation, and summative assessment is conducted at the end of each module. The K-6 program is aligned with the National Science Education Standards, and to science standards in over forty states. Pointers to additional information about the FOSS program are provided in the references under FossInfo (2006). FOSS is used in every state in the United States, with over 100,000 teachers and 2 million students and is in approximately 16% of the nation's school districts. It was the first non-textbook curriculum to make the California adoption list in 1992 and has been adopted again in California for the 2006 science adoption. The program is on 15 state adoption lists, and is used in 50 of the country's 100 largest school districts. FOSS is cited as exemplary in publications by prominent science education organizations, including the National Science Resources Center (Resources for Teaching Elementary School Science, NSRC, 1996), Science for All Children (NSRC, 1997); and the National Science Teachers Association (Pathways to the Science Standards, NSTA, 1996).

Research has demonstrated significant gains in student achievement on standardized science tests (and associated gains in math and literacy) in school districts that use FOSS (Valdez, 2001; Klentschy, 2002). Still, many students, especially in low-performing schools, do not achieve their potential. For example, in the Boulder Valley School District (the site of our study), where all students receive science instruction through FOSS investigations, Colorado CSAP science scores of 1,954 fifth graders in 2006 classified 59.2% as proficient (33.5%) or advanced (25.7%). In the top ten scoring schools, the median number of students in the proficient/advanced category was 80% (range: 73-92). In the ten lowest scoring schools, the median numbers of students in the proficient/advanced category was 36% (range: 5-49), with over 60% of students classified as unsatisfactory or partially proficient.

Embedded Assessments: The FOSS program uses formative assessments embedded in each science investigation to provide teachers (and students) with information that can be used to guide subsequent teaching and learning. According to Malone and Long (2006): "...formative assessment is a critical component of effective instruction: "It is not enough to do activities and to have discussions; you need additional information about how the students are interpreting these activities and discussions" (Black & William, 1998; NRC, 2001). Embedded assessments provide diagnostic information about student learning to both teachers and students as teaching and learning are happening. In FOSS, embedded assessments that accompany each science investigation involve teacher observation (watching students'

inquiry practices during investigations), analyzing and providing feedback to students based on written entries in their science notebooks, and having students engaging in self-reflection and classroom discussions.”

During the past three years, as part of the NSF-supported ASK (Assessing Science Knowledge), new and improved materials and measures for assessing science learning have been developed and tested by Linda De Lucchi, Larry Malone, Kathy Long, Mark Wilson and other members of the FOSS research team in collaboration with classroom teachers (Malone et al., 2004; Malone & Long, 2006). For each science module, the ASK investigators defined the learning goals and objectives of the module in terms of key concepts; what the students should know and be able to do after completing the module. In addition, key concepts were analyzed in terms of their sub-concepts— “the pieces of knowledge that students must know and put together in relationships in order to build the bigger ideas.” Based on this analysis, the ASK project team created construct maps for each module represented as a matrix that describes the key concepts in each science module in grades 3-6, and the constructs that need to be developed to fully understand the key concepts. These construct maps are then used as the basis for developing assessment items to elicit evidence of student learning of the constructs and key concepts in each module. Following Bybee (1996), we refer to the accurate representation of concepts as *conceptual knowledge*, which includes knowledge of vocabulary, definitions and properties of objects, relationships among objects and other concepts that form the foundation for accurate mental models and the basis for scientific reasoning in a specific domain. The ability to use these concepts to explain and predict scientific phenomena using the principles and language of science is called *procedural knowledge*. (My Science Tutor, described in the next section, aims to facilitate learning of conceptual knowledge through guided scaffolded learning in multimedia environments, and learning of procedural knowledge through tutorial dialogs.)

Practical Realities: The Gap between Pedagogy and Practice: Recent studies provide important insights about why some students fail to become proficient in science while using high quality science curricula. Ruiz-Primo et al. (2004) analyzed student entries in science notebooks in two FOSS science modules in six 4th and 5th grade classrooms. Each entry was scored on a four-point scale according to conceptual understanding (defining, exemplifying, relating, comparing or contrasting unit-based concepts) and procedural knowledge (when the communication referred to reporting procedures, observations, results, interpreting results or making conclusions). Their analyses of over 600 science notebooks revealed that notebooks provided valid and reliable estimates of students’ comprehension of science (based on high inter-rater reliability scores and high correlations with other assessment measures). Thus, students’ entries in science notebooks provided valuable information about what students did not know as well as the nature of their missing knowledge, misconceptions and learning challenges.

One of the most striking findings of the study was that teachers in the six classrooms studied *did not provide students with any feedback about their notebook entries*. Given this result, which may be typical of many low-performing classrooms, it is not surprising that formative and summative assessments revealed that many students’ “communication skills and understanding were far away from the maximum score and did not improve over the course of instruction during the school year.” These results suggest that formative assessments that provide accurate and useful information about what

students do and don't know *must be used by the teacher to provide feedback to the student and to guide teaching and learning.*

It is a sad reality that in many classrooms, especially those with a majority of students with limited English vocabulary or prior exposure to science, teachers do not have sufficient time to provide students with the level of individual attention they need. Moreover, many teachers do not receive the professional training required to interpret student behaviors accurately, provide effective feedback and/or adapt their teaching activities to meet individual student's needs. Yet recent studies conducted in the El Centro School District in California have produced promising results and suggest that providing teachers with sufficient training and resources to analyze student performance and provide effective feedback through guided scaffolded instruction can produce significant learning gains. In these studies, which used FOSS, STC or Insight science kits for different science topics, teachers in 4th and 5th grade classrooms received approximately 100 hours of professional development on how to assess and provide scaffolded learning during science investigations. Students who received scaffolded instruction demonstrated moderate to large learning gains on standardized tests of science achievement relative to students who did not receive scaffolded instruction in their science investigations (Klentschy, 2007).

To summarize: 1) Implementation of high quality inquiry-based science curricula such as FOSS produces significant gains in school districts that use these programs, 2) Many students in these districts fail to achieve their potential when using these programs, 3) While formative assessments embedded in science investigations provide valid and useful information about student achievement and learning challenges, this information may not be utilized well (or at all) by teachers to provide feedback or to adapt instruction to individual students, and 4) at least one study suggests that incorporating guided scaffolded instruction into science investigations can improve student achievement.

On the basis of this evidence, we believe that students in low performing classrooms and schools using FOSS, STC and Insight could benefit greatly from a computer intervention that assesses each student's conceptual and procedural knowledge and provides expert, contextualized, individualized, adaptive instruction in optimal ways. We hypothesize that guided scaffolded instruction incorporating continuous assessment in multimedia environments can be used to provide the conceptual knowledge base necessary to enable students to engage in tutorial dialogs that will guide them to think and reason using the language, methods and principles of science.

Theoretical and Empirical Rationale

This proposal builds on theory and research in areas of (a) human learning, (b) intelligent tutoring systems and social agency (c) human computer interface design and (d) multimedia learning.

Theory and Research on Learning: Cognitive learning theorists generally agree that learning occurs most effectively when students are actively engaged in critical thinking and reasoning processes that cause new information to be integrated with prior knowledge. Discourse Comprehension theory (Kintsch, 1998) provides a strong theoretical framework for asking questions and designing activities that stimulate thinking and construction of deep knowledge that is useful and transferable; this theory

provides the foundation for several instructional approaches to comprehension (King, 1991; Beck et al., 1996; McKeown & Beck, 1999).

Comprehension theory suggests two considerations that are of prime importance here. First, there is the notion of levels of understanding, varying from superficial to deep understanding. Our knowledge of the conditions that foster different levels of understanding can guide instruction. Deep learning requires integration of prior knowledge with new information and results in the ability to use this information constructively in new contexts (the formation of a *situation model*, Kintsch, 1998). Second, the type of mental representation students form (superficial understanding versus adequate situation model) is determined by comprehension strategies students use (Paris et al., 1991; Pressley & McCormick, 1995).

Benefits of Tutorial Instruction: The proposed learning tools incorporate one-on-one interaction with a lifelike computer character or virtual tutor. The tutor guides learning of vocabulary and concepts in exercises that incorporate various media (text, images, narrated animations) into question and answer dialogs that continuously assess the student's knowledge and present effective feedback to scaffold learning. Once conceptual knowledge is mastered, tutorial dialogs are initiated, based on principles used in the Questioning the Author approach (below), designed to stimulate deeper thinking and scientific reasoning that is evidenced by accurate explanations and predictions when presented with new problems.

Theory and research provide strong guidelines for designing effective tutoring dialogs. One source of evidence comes from studies of how effective teachers teach (e.g., Bransford, 1993; Bransford et al. 1990, 1999; Allington, 2001). In their review of how expert teachers differ from novice teachers, Bransford et al. (1999) note that expert teachers have expertise both in their content domain and in their pedagogical content knowledge, i.e., specific teaching strategies that differ from one discipline to another. "Expert teachers know the kinds of difficulties that students are likely to face; they know how to tap into students' existing knowledge in order to make new information meaningful, and they know how to assess their students' knowledge." The effect of the teacher is highly significant. Both Ferguson (1991) and Snow (1989) found that the quality of the teaching is the most powerful predictor of student achievement.

Over two decades of research have demonstrated that learning is most effective when students receive individualized instruction in small groups or one-on-one tutoring. Benjamin Bloom (1984) determined that the difference in learning gains between students who received classroom instruction and those who received either one-on-one or small group tutoring was 2 standard deviations. In the two decades since Bloom's initial research, evidence that tutoring works has been obtained from dozens of well designed studies, meta-analyses of research studies (e.g., Cohen, Kulik & Kulik, 1982), and positive outcomes obtained in large-scale tutoring programs in Britain (e.g., Topping & Whitley, 1990) and the U.S. (Madden & Slavin, 1989). Recent studies and literature reviews (e.g., America Reads, 1997; Chi et al., 2001, Cromley & Azevedo, 2005) provide insights about factors that improve tutoring outcomes, including frequent, regular and well structured tutoring sessions, coordination with classroom activities, well trained tutors, and dialogs structured to stimulate thinking and knowledge construction (America Reads, 1997).

Questioning the Author: The process of dialog design will be informed by principles of dialog interaction incorporated in Questioning the Author (QtA), an approach to classroom instruction developed by Isabel Beck and Margaret McKeown (Beck et al., 1996; McKeown and Beck, 1999, McKeown et al., 1999). We selected QtA as the basis for tutorial dialogs because it is a highly regarded, scientifically-based program that uses dialog interaction to facilitate development of effective comprehension strategies and deep learning. QtA is a mature program developed during the past decade that has been implemented in dozens of school districts across the country and is used by hundreds of teachers (Beck et al., 1996; McKeown and Beck, 1999). The program has well established procedures for training teachers to interact with students, for observing teachers in classrooms and for providing feedback to teachers.

Questioning the Author is a deceptively simple approach with a minimum of apparatus. Its focus is to have students grapple with and reflect on what an author is trying to say in order to build a representation from it. In science investigations, the perspective of the “author” in the context of “Questioning the Author” moves from questions about what a specific author is trying to communicate, to questions about the investigations and outcomes. The “author” may be the observations and data sets that accrue from the active investigations, with open-ended questions leading to dialogs that encourage the student to make sense out of their hands-on experiences, and guiding that sense making to converge on conventional models and understandings of science. Open-ended questions would include: What's going on here? What's that all about? What does that mean? Why is that important? How does that connect to...?

Because the dialog modeling used in QtA is well understood and can be taught to others (Beck & McKeown, 2006), it is well suited for implementation in an intelligent tutoring system. QtA provides a good basis for tutorial interaction in the proposed virtual tutoring system because (a) research shows that it is effective for improving comprehension (Murphy & Edwards, 2005), (b) it provides a framework and planning process that helps define learning goals and develops an orderly sequence for getting students to achieve the goals, (c) it offers ways to design prompts that draw student attention to relevant portions of presented material, but that are open enough to leave the identification of the material to students, (d) it provides a principled, easily understandable and well documented program for teachers to elicit and respond to student responses that helps them learn to focus on and make connections between meaningful elements of the discourse and their own experiences, and (e) it focuses on comprehension, with discussion of student personal views and experiences limited to those that can directly enhance building meaning from texts, lectures, multimedia presentations, data sets, or hands-on learning activities.

A recent IES funded study conducted by Ian Wilkinson and colleagues examined approaches to conducting discussions with students, QtA was identified as one of two approaches out of the nine examined that is likely to promote high-level thinking and comprehension of text. Relative to control conditions, QtA showed effect sizes of .63 on measures of text comprehension, and of 2.5 on researcher-developed measures of critical thinking/reasoning (Murphy & Edwards, 2005). Moreover, analysis of the QtA discourse showed a relatively high incidence of authentic questions, uptake, and teacher questions that promoted high-level thinking—all indicators of productive discussions likely to promote learning and comprehension of text (Soter & Rudge, 2005). We note that, while some teachers

have extended QtA dialogues to science topics, QtA has been implemented mainly in the language arts environment; an innovative advance of our work is to apply the tested and true methodology to a new domain, science.

Benefits of Intelligent Tutoring Systems: My Science Tutor is one of a class of systems that are designed to enhance student achievement by providing students with individualized instruction similar to that provided by a knowledgeable and effective human tutor. Effective intelligent tutoring systems exist today in the laboratory and a few have been deployed in real-world applications. These systems typically use natural, text-based dialog interaction; some incorporate an animated pedagogical agent. Experiments with scientifically-based intelligent tutoring systems have demonstrated up to one sigma learning gains (approximately equivalent to an improvement of one letter grade, e.g., from C+ to B+) when comparing performance of high school and college students who use the tutoring systems to students who receive classroom instruction (Graesser et al., 2001; Van Lehn & Graesser, 2001; Van Lehn et al., 2002).

My Science Tutor will be distinguished from most prior systems by a combination of factors: inclusion of a lifelike 3-D character (a pedagogical agent or virtual tutor) that produces accurate visual speech and emotions synchronized with naturally recorded speech, the integration of narrated multimedia animations within dialogs to optimize acquisition of conceptual knowledge, the application of both domain-dependent and corpus-based domain-independent shallow semantic parsing approaches in tutorial dialogs, and the use of an effective and well-tested approach to dialog design, *Questioning the Author*.

Benefits of Pedagogical Agents: A significant body of research, stimulated by the seminal work of Reeves and Nass (1996), has demonstrated that people interact with computers using the same fundamental rules and social conventions that guide our interactions with other people. When human computer interfaces are based on the social conventions and expectations that govern our daily interactions with each other, they provide engaging, satisfying and effective user experiences (Johnson et al., 2000; Reeves and Nass, 1996; Nass & Brave, 2005). Such programs foster *social agency*, enabling users to interact with the program like they interact with people. Programs that incorporate pedagogical agents, represented by talking heads or human voices, especially inspire social agency in interactive media (Atkinson, 2002; Baylor et al., 2003, 2005; Mayer, 2001; Moreno et al., 2001, Nass & Brave, 2005; Reeves and Nass, 1996). In comparisons of programs with and without talking heads or human voices, several studies have reported that subjects learned more and reported greater satisfaction using programs that incorporated virtual humans (e.g., Moreno, 2001, Atkinson, 2002; Baylor et al., 2003, 2005).

Anticipated Products

Our project will produce computer software to supplement and improve science achievement by 4th and 5th grade students using FOSS in four areas of science. The software is intended to facilitate learning of both conceptual knowledge and procedural knowledge. While the software is linked to the FOSS program, it will be designed to provide a sequence of well organized and self contained tutoring sessions that can be used independently by students, and should facilitate learning of vocabulary, concepts and scientific principles in specific areas of science if used independently of FOSS or other programs. By

designing self contained learning tools linked to FOSS science investigations, the program both complements and supplements the FOSS program.

In addition to providing learning tools for teachers and students, we will distribute a free toolkit that researchers can use to modify and conduct experiments with our applications and to research and develop new applications that incorporate tutorial dialogs with virtual humans in multimedia environments.

WORK PLAN

For ease of exposition, we describe the research and development effort in terms of activities that facilitate learning of *conceptual knowledge* and activities that facilitate learning of *procedural knowledge*. During a typical 20 minute session in My Science Tutor, we expect students to spend the first 10 minutes learning key concepts and the second 10 minutes in tutorial dialogs that stimulate thinking and reasoning about them. In practice the dialog model will be designed to detect incomplete knowledge or persistent misconceptions. When this occurs, the tutor may provide an explanation or present a narrated animation, or the program may exit the dialog to return to learning concepts. In the remainder of this section, we describe the two program components in more detail, and the proposed research and development activities.

Learning Concepts in My Science Tutor

The teaching and learning of conceptual knowledge in My Science Tutor is designed to help students construct accurate mental models that form the basis for knowledge discovery. An accurate mental model assigns meaning and relevance to entities in terms of their properties and behaviors and the rules that govern their interactions. An accurate mental model is sufficiently *coherent and elaborated* to support thinking and reasoning about the scientific concepts within a domain so the student can learn to explain observations or make predictions. Because the rules of science often contradict student's intuitions about how the world works based on everyday experiences, a critical component of the research and development process is to design questions and response choices to elicit preconceptions and misconceptions, and provide feedback and multimedia presentations to enable students to construct accurate mental models.

Gee (2004) has argued eloquently that well-designed virtual worlds (video games) provide a coherent context for situated learning in which children and young adults construct sophisticated mental models that enable them to reside and excel within the virtual world. It is remarkable that tens or hundreds of thousands of students who are failing to learn science in classrooms have demonstrated exceptional learning and achievement in virtual worlds with concepts and rules more complex than those presented in classroom science lessons. My Science Tutor attempts to provide a virtual world that contextualizes and compliments classroom science investigations to make learning more relevant and meaningful.

My Science Tutor will teach vocabulary and associated concepts through contextualized and scaffolded learning activities. An adaptive study plan tracks all system presentations and student responses to questions, and sequences activities based on the student's responses. Correct responses cause the tutor to optionally provide positive verbal or nonverbal feedback and/or expand on the correct answer.

The student is then presented with a new question, which may assess knowledge of the concept from a different perspective. Students who answer all or most questions correctly move quickly through the study plan and proceed to spoken dialogs. “Thinking Multiple Choice” questions are designed to assess common confusions and preconceptions that are inconsistent with the concepts of science being studied.

When a student makes an incorrect response, the tutor will typically provide a hint and ask the student to make another choice. After an incorrect second choice, the tutor presents the correct answer. The correct answer may be followed by a spoken explanation accompanied by a sequence of illustrations or an animation that describes the concept. Periodically, the student’s retention of vocabulary and concepts is reviewed. When the student’s responses to multiple choice questions suggests that they have constructed accurate mental models of key concepts, My Science Tutor engages the student in tutorial dialogs in which students must use the concepts they have learned to explain and predict scientific phenomena. A typical learning session at the beginning of a science investigation will begin with the virtual tutor providing a brief introduction and overview of the science topic. Next, the tutor will introduce important concepts through vocabulary instruction, with short narrated animations that help the student visualize and comprehend concepts associated with key words or phrases (e.g., attraction, repulsion, condensation, evaporation). The student is then engaged in a variety of exercises that continuously assess and train comprehension of the meanings and uses of the words and phrases as they apply to the concepts and practice of science.

Once the set of vocabulary items associated with a set of science concepts have been mastered, a narrated animation is presented that explains and shows the main concepts that the student is expected to learn in the FOSS science investigation, and relates them to both of the science experiments or common and relevant real-world situations. The initial narration will be designed to contextualize learning of science in a specific area (e.g., Mixtures and Solutions, Landforms) by presenting the main concepts in an engaging and entertaining way using vocabulary that the student has learned.

Software Development: During the conceptualization and planning process, the design team, in collaboration with the educational researchers and teachers, will brainstorm and converge on a set of animated narrations for teaching key concepts. The design of narrated animations and multiple choice assessment questions will be closely linked to the learning objectives, construct maps and embedded assessment questions resulting from the FOSS ASK project. The design of response choices and of narrated animations that follow incorrect response choices will be informed by analysis of transcriptions of notebook entries that indicate students’ most common confusions and misconceptions.

My Science Tutor assesses learning and comprehension using *thinking multiple choice (MC) questions* based on principles developed by our research team as part of the Colorado Literacy Tutor (Wise et al., 2003; Kintsch E., 2005). Thinking MC questions occur at logical points within a story and following the story. The design of thinking MC questions was motivated by the Discourse Comprehension theory and informed by research conducted by Beck, McKeown and others (1996). The goal of thinking MC questions is to stimulate the student’s thinking (a) through thought provoking questions, (b) by providing response choices that challenge the reader and assess the nature of comprehension

difficulties, and (c) through feedback on response alternatives that stimulate additional thinking and deepen comprehension.

During the development phase, and perhaps also in the final system, students will be prompted to produce spoken answers to each question before being presented with multiple response choices. We will transcribe, read and cluster similar answers to identify common confusions and misconceptions. We will combine the results of this analysis with confusions and misconceptions identified by FOSS researchers, and use the knowledge gained to design and refine response choices and feedback to incorrect choices. We predict that having students produce spoken answers before reviewing response choices may facilitate learning, based on research by Chi (1996), Van Lehn et al (2003) indicating that science learning improves when students are able to verbalize their beliefs about scientific phenomena. We plan to conduct experiments in year 3 comparing student experiences and learning outcomes when students do and do not produce spoken responses before choosing an answer from multiple choices.

Infrastructure and Expertise for Concept Learning in My Science Tutor: Research and development of My Science Tutor leverages significant infrastructure and prior work by members of the project team. Since 2000, with support from an NSF ITR grant, Cole and his colleagues at CSLR have developed six programs that use virtual tutors and therapists; each of these programs has undergone summative evaluation of clinical trials, with positive outcomes. Descriptions of these programs, in which virtual tutors assess or teach reading, or virtual clinicians conduct speech and language treatments with individuals with aphasia or Parkinson disease, are presented in Cole et al. (2007a, 2007b), van Vuuren (2007), and on the CSLR Web Site (VH-Web, 2007). These programs provide significant infrastructure for My Science Tutor, including the software architecture, design tools, adaptive study plan, narrated animations and integration of speech recognition and 3-D character animation systems enabling natural conversational interaction with Marni, a lifelike pedagogical agent that produces accurate visual speech and emotions. The project team also has significant expertise developing engaging and effective learning tools for elementary school students. Foundations to Literacy (Cole et. al., 2007a), which uses guided scaffolded learning to teach foundational and fluent reading skills to K-3 students through conversational interaction with Marni, has demonstrated significant learning gains in letter and word recognition and comprehension in summative assessments with over 2000 children in 50 classrooms. Interviews of over 500 students and 30 teachers produced very positive ratings; both students and teachers believed the program was engaging and effective.

Tutorial Dialogs

The goal of the dialog development effort is to create tutorial dialogs with virtual tutors that are as engaging and effective as spoken dialogs with human tutors. In this section we describe the process for developing accurate and natural tutorial dialogs based on a proven approach, QtA, to improve science learning in FOSS science investigations.

During the initial phase of the project, a team of 6 QtA project staff (Ward, Cole, van Vuuren, former teachers (project tutors) and graduate students) will be trained in QtA dialog techniques by Dr. McKeown in consultation with Dr. Samantha Messier, BVSD Science Coordinator, FOSS co-developers Larry Malone and Linda De Lucchi and educational math and science researchers Eric Hamilton and

Barry Sloane. Training will begin with a three-day workshop. Participants will learn about the theoretical and empirical background of QtA, view videos and analyze transcripts of master QtA teachers, take part in QtA dialogs and conduct practice tutoring sessions that will then be the focus of group discussion, including feedback by Dr. McKeown. Following the participants' trial lessons, Dr. McKeown will provide a demonstration lesson in one of the science content areas, which will then be discussed by the group. Participants will then plan a QtA lesson, in collaboration with the science educators and researchers after completing and studying the materials that accompany a FOSS science investigation. Participants will try out their lessons with students.

Following the introductory workshop, the trained project tutors will work with Drs. Malone, De Lucchi and Messier to develop and conduct tutoring sessions for content in FOSS science investigations. These sessions will be videotaped and transcribed, and the transcripts analyzed by Dr. McKeown and her colleagues. Feedback to tutors and project staff will be provided. A cycle of lessons and feedback will be set up so that novice tutors do a lesson, receive feedback, incorporate the feedback into their tutoring, and then provide another video transcript for critique. Three rounds of feedback are planned during year 1.

Developing automated tutorial dialogs: When a tutorial dialog is created, we specify each of the key concepts that the student is expected to know based on the content, instructional activities and learning objectives of the associated FOSS science investigation. In the completed dialog, the system initiates the dialog, based on QtA principles, by asking questions that assess the student's knowledge of a key concept. Following QtA guidelines, the segment begins with an open question (referred to as a query in QtA), which asks the student to relay the major ideas presented in a narrated animation or a classroom science investigation. Follow-up queries draw out important elements of the investigation that the student has not included. The follow-up queries are created by taking a relevant part of the student's response and asking for elaboration, explanation, or connections to other ideas. Thus the follow-ups focus student thinking on the key ideas that have been drawn from the investigation. If the student is unable to construct adequate explanatory representations of ideas from an investigation, portions of the lesson are explained (using optional illustrations and animations) and queried to refocus student thinking. Each question the tutor asks has a set of points (e.g., vocabulary words, concepts, relationships, associations) that the answer should contain. The components of the dialog system (speech recognizer, semantic parser) analyze the utterances produced by a student in response to the question, and determine what points have been presented in the answer, and what points are not within the student's response. Questions are generated about the missing or erroneous concepts to attempt to elicit information about them.

Following the initial phase of development, the system is field tested with students, first in the laboratory (years 1 & 2) and then in classrooms (years 2 & 3) in a set of schools selected for the development phase of the project. The initial testing of dialogs between the student and the virtual tutor will be monitored by project tutors who act as invisible "Wizards" in QtA dialogs who can override the system's dialog responses in order to present the student with a correct response if the system makes an error. These so-called Wizard of Oz experiments provide an important first-hand mechanism for observing, analyzing and evaluating student interactions with the virtual tutor. The Wizards (project

tutors), who have been trained to be expert Q&A tutors for each science investigation, make judgments about each response that the system is about to produce during the dialog session. If the Wizard validates the staged response, the system delivers the programmed response to the student. If the staged response is not validated by the Wizard, the system delivers a revised response (with synthetic speech) that the Wizard types in on the spot. In either case, the Wizard is providing real time evaluation of the dialog, and producing a real time assessment that the development team can analyze, discuss and use to improve dialogs. The data collected during these sessions is used to retrain the speech recognizer's acoustic and language models, to test and improve coverage of the semantic parser, and to test and refine the dialog model.

Following approximately 6 months of dialog development for the set of science investigations in each FOSS module, the dialogs are field tested in classrooms with individual students. The dialog testing, observation and recording, transcription, evaluation, refinement and re-testing process will continue in classrooms throughout the second and third years of the project. During initial classroom field testing, a project tutor will observe dialog sessions. Depending upon the performance of the system at this point, the tutor may intervene as a Wizard who takes control of the program if the dialog stalls. Data collected during field tests will be used to improve the program through a series of iterative design and test cycles.

Architecture of the Dialog System: The proposed system for dialog interaction relies on a domain-specific shallow semantic representation. Shallow semantics refers to making explicit the underlying predicate argument structure, which can be characterized as a semantic Frame together with its Frame Elements, the semantic slots and fillers which are the predicate arguments. This mechanism represents the entities, events and concepts in a domain and it also represents the relations between them. The representation is shallow enough to be robust and efficient and specific enough to be precise and useful. For each concept to be discussed, the module developer must create one or more frames that represent the concept and its elements; all of the events, entities and relations that are relevant. As an example, the frame for the concept "vibrations produce sound" appears on the right.

Each predicate and frame element is associated with a semantic grammar that matches all semantically equivalent paraphrases such as "vibrating objects make noise". The system gains much of its robustness from using grammars to model word strings that correspond to a frame element (slot) rather than sentences. Frame elements tend to correspond to relatively short phrase level strings, so it is easier to get good coverage of the sequences than for sentence level strings. Semantic grammars are used to model the frame elements rather than standard syntactic grammars. This provides a simple mechanism for specifying the many different, semantically equivalent, word strings that would specify a concept. This not only allows for synonyms, but also for idioms and common metaphors, and is crucial to the flexibility of working reliably with youngsters of varying language abilities.

<p>Frame: Vibrations Produce Sound</p> <p>Predicate: Produce</p>

Resources and Expertise for Dialog Development The expertise needed to develop tutorial dialogs has been developed and demonstrated by the project team assembled for this project. In the DARPA

Communicator program (Pellom et al., 2003), which supported research and development of natural spoken dialog systems, a team led by Dr. Ward and Dr. Cole developed tools and technologies that formed the Conversational Agent Toolkit (Ward & Pellom, 1999; Pellom, et al., 2000). The CU Communicator system supported natural mixed-initiative dialogs by telephone callers who made travel plans that included air travel, hotel reservations and car rentals. Independent assessment of the CU Communicator system resulted in 95% call completions (with an itinerary emailed to the caller) and the highest user satisfaction rating achieved by any spoken dialog system in the program. This success in creating mixed initiative dialog systems is an important forerunner to the efforts proposed here to develop dialog tools for science learning. In mixed-initiative dialogs such as these, either the user or the system can seize the initiative and take control the dialog. This is a critical feature for effective tutorial dialogs. For example, the tutor may ask the student a question, and the student may respond by asking the tutor another question. Well designed mixed-initiative dialogs are able to handle such flexible conversations in a graceful and natural way. In addition to not requiring that the user respond to a question asked by the system, these systems are able to deal with very open ended questions and often initiate a conversation with a very general question such as “How may I help you?” This capability is a very good match to the requirements of QtA style dialogs.

Pilot Experiment: We conducted a pilot experiment to determine how well an automatic system based on domain-specific shallow semantic parsing could grade summaries of stories. Summaries were collected after students in grades 3-5 listened to a short story, and were asked to tell the experimenter “what the story was about.” While summaries are different than explanations given in QtA dialogs, they provide a good test bed for investigating the ability of natural language processing techniques to assess students’ comprehension of the main ideas presented in a text. For the pilot study, we analyzed summaries of one story, “Racer the Dog” spoken by third and fourth grade students. There were 22 student summaries for the story. The summaries were divided into a training set of 15 and a test set of 7. A reference summary was generated by the experimenters that contained the points that should be contained in a summary. The Phoenix semantic parser from the CAT toolkit was used to map summaries to semantic frames. This is the same tool we will be using in the proposed project. A grammar was written for Phoenix that could extract the points in the reference summary. Anaphoric reference was resolved manually in a pre-processing step. Then the 15 training summaries were used to expand the coverage of the grammar. The test summaries were not looked at for developing the system. The 7 test summaries were then parsed by the system. The test set was also parsed manually to create reference parses so that the accuracy of the parser in extracting the points could be measured. There were 36 points in the hand parsed test summaries. Compared to these, the automatic parses had a Recall of 97% (35/36) and a Precision of 100% (35/35). The parser found all but one of the relevant points and produced no erroneous ones. In order to be considered correct, the concept from the summary must have the same semantic roles as one in the reference, and the roles must be filled by the same entities (or references to the same entities).

We decoded the speech files with the University of Colorado SONIC speech recognition system (Pellom, 2001; Pellom and Hacioglu 2003) and processed the SONIC output instead of human generated transcripts. For the same test set, the results were: Recall= $30/36= 83\%$ and Precision= $30/30= 100\%$.

Speech recognition errors caused an additional 5 points not to be extracted, but generated no erroneous ones. Tutorial dialogs are designed to seek clarification and use other conversational conventions to maintain natural and graceful dialog interaction when information is missing from a student's response regardless of whether the information was missing in the student's response or because of a system error.

RESEARCH TO ENHANCE GENERALITY AND PORTABILITY OF DIALOGS

The process of extracting information from student answers and comparing it to information in reference frames is a form of **entailment**. There are sets of reference propositions that the student should know. Entailment is the process of matching the propositions in the students' statements against the reference set of propositions and determining if each of the reference propositions is entailed, i.e., it can be inferred from the student's answer. We are using the term "proposition" here very loosely to mean a specific relation between entities, events and concepts, such as "vibrations produce sound".

Dialogs are oriented toward asking follow-on questions about the reference items that have not yet been entailed. Much of the effort of creating dialogs for modules is enabling the entailment with a robust domain specific representation. One of the most expensive and labor intensive parts of producing dialogs for a new module is the development of the semantic grammars for the frame elements that specify all semantically equivalent paraphrases. The tool is domain independent, but it operates on a domain specific representation. Using domain independent representations for entailment would greatly reduce the effort of producing new modules. One of the most important research and development features of this project is the significant technical progress we have made in this direction. Pilot experiments have indicated that the dependency parser and thematic role labeler used by the system are robust with respect to the type of ungrammatical sentences produced by children. The data used were answers from fourth graders to the benchmark questions from a number of the FOSS modules for fourth grade.

Field Testing My Science Tutor

Field testing in classrooms is an integral part of the development process. The idea is to develop an integrated set of My Science Tutor learning activities as soon as they are ready for field testing for each of the 4-6 science investigations in the 4 FOSS science modules used in our study. This requires efficient scheduling of new software releases to coordinate with science instruction underway in Boulder Valley classrooms. Fortunately, all FOSS science modules are taught simultaneously throughout the school year, so some schools classrooms will always be available to conduct field tests of our learning tools. (The attached letter of support reveals Boulder Valley School District's (BVSD's) commitment to support our project and to facilitate scheduling of classrooms for field testing and assessment. The project's subcontract to BVSD will pay a half time staff person, hired through the district, to schedule and coordinate field testing and assessment activities.) During field testing, students in classrooms and computer labs will be rotated onto the computer for 20 minute sessions. Project tutors will be present full time during the first few days the program is used in each school and will observe students at other times during 20-30% of their sessions. The observers will note whether the student is able to use the program without human mentoring, if the student appears to be engaged and productive using the

program, and if the student experiences problems or confusion during program use. The observer will intervene if the student asks for help or is obviously stuck. These observations will be used to improve human-computer interaction within the program. We will also work closely with interested teachers to help them learn about the program, and to establish effective communication protocols for responding to problems quickly whenever they occur, and coordinate new releases with their teaching plans.

A portion of students who use the program will be interviewed to assess their experiences and opinions and solicit suggestions for improvements. We will analyze and report the results of the usability study by documenting time on task for all students, documenting and reporting all problems that occur, analyzing and presenting results of observations, student and teacher interviews and by measuring students' progress within the program's study plan. These measures will help us to identify design problems and propose solutions to these problems that will lead to more engaging and effective experiences. We will field test My Science Tutor with 40 to 60 children for each FOSS science module.

EVALUATION

The assessment plan will be implemented during year 4 and in year 5. Year 5 will serve as a replication of the study with refinements to the software and implementation of the plan based on year 4 experiences. Independent evaluation will be conducted by Dr. Tim Weston and his team, with collaboration and oversight by Finbarr (Barry) Sloane on planning and analysis of results. The objective of the Phase II evaluation effort is to compare gains in science learning and reading comprehension among students who use My Science Professor to students in classroom control conditions. Students randomly assigned to the treatment and control conditions will receive the same amount of time receiving supplementary science instruction following FOSS science investigations conducted in classrooms.

Summary: Summative evaluation will compare learning behaviors and learning gains of students in fourth and fifth grade classrooms randomly assigned to one of three conditions: (a) My Science Tutor, (b) Human Tutoring, and (c) classroom instructional comparison. Immediately following classroom science investigations in the FOSS modules entitled **Landforms, Water, Variables and Mixtures and Solutions**, students in the study will receive 20 minutes of additional instruction in one of three conditions:

1. Computer Treatment: Students in the My Science Tutor condition will leave the classroom to receive approximately 20 minutes of individualized instruction using My Science Tutor in a computer laboratory.
2. Human Tutoring: Students in the human tutoring condition will leave the classroom and receive small group instruction with a project tutor in a quiet environment. We chose to compare learning outcomes for human tutoring in small groups (averaging 3 students) rather than one-on-one tutoring because learning gains with human tutors have been found to be equivalent following one-on-one and small group instruction sessions (e.g., Bloom, 1984; REF). Given this result, it is logistically easier and less expensive to use three tutors each with three students than nine tutors each tutoring a single student.
3. Classroom Control: Students in the control condition will remain in the classroom and read and possibly discuss stories that are included with each FOSS module as supplementary reading materials.

Hypotheses

Based on prior research and our own expectations, we predict that summative assessments will produce the following pattern of results:

1. Students in the Computer Treatment condition will demonstrate significant learning gains in science on FOSS benchmark tests and Colorado CSAP tests for relevant science items relative to students in the classroom control condition. We expect to observe moderate learning gains, with effective sizes between .4 and .6 based research cited above by Mayer, Van Lehn and others using intelligent tutoring systems. While larger effect sizes have been obtained with intelligent tutoring systems compared to classroom control comparisons, it is important to remember that FOSS is a high quality program, so gains relative to FOSS instruction will be an a significant achievement.
2. Students in the Human Tutoring condition will demonstrate significant learning gains in science relative to students in the classroom comparison condition with moderate to large effect sizes of .6 to 1.0. We expect human tutors to be more effective than virtual tutors relative to classroom controls, since computer programs cannot yet match the dialog capabilities and judgments of expert tutors.

Formative and Summative Evaluation

FOSS Embedded Assessments: Student entries in science notebooks provide an excellent opportunity to observe learning differences across treatments. Based on the work of Ruiz-Primo et al. (2004) described above, we assume that student entries in science notebooks can be reliably scored for learning of conceptual and procedure knowledge during science investigations. Independent assessors blind to the student condition assignment will score notebooks. We will analyze learning gains of students within science investigations in each condition. We hypothesize that students will benefit from concept learning activities in My Science Tutor and their notebook entries will reveal fewer confusions and misconceptions about key ideas. We also expect that notebook entries will provide evidence of improved learning over the course of science investigations relative to students in the classroom comparison condition. To understand effects of classroom instruction during science investigations, we will observe and code classroom instruction interactions, including types of questions asked by teachers and students and responses to same and correlate learning outcomes to coded observations of classroom instruction.

Members of the project staff will observe students in the human tutoring treatment to assure fidelity of treatment and to make notes on social interactions and instructional practices in the human tutoring condition. About 20% of the sessions will be video taped for subsequent analysis. Fidelity of treatment in the computer treatment condition is guaranteed by the computer program's study plan, which assures consistent presentation of materials and predictable interactions between the program and the student based on analysis of their prior responses within the program. Following the intervention, feedback will be gathered from both students and teachers with surveys and interviews designed and field tested for this purpose. All qualitative data will be analyzed with a domain analysis (Spradley, 1980).

FOSS Benchmark Assessments: Summative evaluation of science learning in our study will use (a) the FOSS ASK summative evaluation measures developed for the four FOSS modules used in our study, and (b) analysis of relevant science items on the Colorado CSAP test administered to fifth grade students.

Independent Evaluation will be conducted by ATLAS Center at the University of Colorado under the direction of Dr. Timothy Weston. Pre-tests, post tests and delayed post-testing will include content to be taught in the FOSS modules. These benchmark tests have been developed by Mark Wilson at UC Berkeley as part of the ASK project. In addition, the psychometric group at UC Berkeley has also developed embedded assessment measures for each investigation (i.e. each two week unit). The benchmark assessments have between 8 and 12 items and show composite reliability with alphas in the .80' and 90's. The interrater reliability for subjective items also meets high standards, and the validity of the measures has been built up over time through a process of empirical investigation.

Colorado State CSAP Measures: In addition to the measures described above we add one more: The Colorado CSAP state wide test of science administered to all students in grade five. Consequently, we will also evaluate the effects of the tutoring program on differential student performance by assigned group.

Student Characteristics and Power Analysis

A maximum of 216 students in each grade (total of 432) will participate in the study during assessments conducted in years 4 and 5. Each student will receive instruction in only one FOSS module during a ten week period and will receive supplementary instruction in one of the three conditions. These students will be drawn from fourth and fifth grades in Boulder Valley schools that scored in the lowest one third on Colorado CSAP science tests administered to students. Four different schools (16 classrooms) will participate in the study; for logistical reasons, a participating school will have two fourth and two fifth grade classrooms, enabling assessment to be conducted for $27 \times 4 = 108$ students in each school, or 36 in each condition. Thus one school, or 108 students, will participate in the study for each FOSS module. The four schools participating in the study have approximately 50% of all students from lower SES backgrounds and a higher proportion of Hispanic students and English language learners than higher performing schools.

Assignment and Power: Students will be randomly assigned within each of the 16 classrooms to one of three conditions for a period of ten weeks. Before the beginning of the study demographic information from all students will be gathered to provide the basis for randomly assigning within gender and language status blocks, and as a check for differential mortality in case of student attrition or non-participation in the study.

We conducted our power analyses using conservative estimates of effects while also utilizing a one-factor Analysis of Variance Design that allows effect sizes to vary between conditions. Students experiencing human and computerized tutors are expected to differ significantly in outcomes between these two conditions ($ES = .2$). A moderate effect size difference of .4 of a deviation unit or more is expected between the computer treatment and classroom instructional control groups. Given these expectations, the power for a three group comparison with 8 classrooms of 27 students each ($n = 216$) provides more than 80% power at the .05 level, with attrition rates of 20% over the ten week intervention period. Power is approximately 65% for the alpha .01 level.

Research setting, confounds, treatment fidelity

Because of the relatively short (10 week) duration of the treatment and the fact that the students in the treatment and control groups are separated physically from each other, risks of treatment contamination between students is minimal. Physically removing students from the classroom to receive the computer-based treatment is also expected to minimize compensatory response to the intervention by teachers or students in the control condition. Scores from pre and post measures and follow-up measures will be used as dependent variables for analysis. The three FOSS module assessments will be treated as equated measures and module type will be used as an independent variable in the analysis. We will incorporate hierarchical modeling to learn the intra-class correlation and the percentage of variance due to class membership. Additionally, the pre, post and follow-up measures will be analyzed with HLM growth modeling tools (Raudenbush & Bryk, 2002). We do this because of the flexibility of these newer models to deal with missing data, and uneven spacing of data collection opportunities. When the timing of data collection is spaced equally and when there are no missing data these growth curve techniques provide equivalent results to the Repeated Measures Analysis of Variance (RANOVA). However, this is rarely the case in non-laboratory settings and HLM techniques provide us more flexibility from an analytic perspective. In sum, these techniques allow us to better maintain our sample size over time, allowing for the inclusion and analysis of students with incomplete data. Other variables of interest incorporated into models may include person characteristics such as gender, ethnicity and language status, and differences between schools. All statistical analyses will include effect sizes, and where appropriate clustered effect size.

Timelines, Milestones and Dissemination

The project will be managed by PI Wayne Ward in close collaboration with co-PIs Ron Cole and Sarel van Vuuren. This team has substantial experience managing software development projects and assessing system performance and learning outcomes. We took care to incorporate timelines and milestones related to the development process within the work plan section. Timelines and milestones are briefly reviewed here for the entire project. Year 1 will focus on provisioning BLT, hiring staff, planning timelines and milestones and associated management and staff responsibilities in detail, organizing an all-hands planning meeting, conducting QtA workshops and training, specifying the system architecture and software development plan, initiating software development activities, and developing and testing infrastructure for Wizard of Oz experiments. Year 2 will focus on staff training and coordinating, implementing and refining human tutoring activities for 2 FOSS modules, software development of these modules, and coordinating human tutoring, software development and Wizard of Oz studies. Year 3 will continue iterative design and test of the software for the first two FOSS modules and implement field testing of same in the first school semester. The remaining 2 FOSS modules will be developed for field testing in the second semester. In Year 4 My Science Tutor will be assessed. Developers will improve the program based on analyses of system performance and formative and summative assessments. The data collected during the assessment will be analyzed and reported. In Year 5, formative and summative assessments will be conducted again using the improved software. The program will be integrated into the Virtual Human Toolkit for free distribution to university researchers. Results will be analyzed and manuscripts posted on the project Web site and submitted for publication.

Dissemination. The results of this project will be communicated to STEM professionals and practitioners during the project via a project Web site that will be set up and maintained by BLT during the first year of the project. We will also present results of the software architecture and applications, field testing outcomes and assessment plan and outcomes at national and international conferences in several fields that are attended by the computer scientists, engineers, cognitive scientists, mathematicians and educational researchers who comprise the project team. During the 3rd – 5th years of the project, we will write and submit articles to leading peer review journals.

A set of software applications resulting from the project will be incorporated into the Virtual Human Toolkit, an authoring environment for designing spoken dialogs with virtual tutors in multimedia environments, and freely distributed via the project Web site to university researchers for non-commercial use. If successful outcomes are realized, we expect to work with the FOSS co-developers and DELTA publishing to prepare the software for commercialization.

RESULTS FROM PRIOR NSF SUPPORT

The goal of NSF-IIS-0325646 Domain-Independent Semantic Interpretation (**Wayne Ward**, PI) was to improve the technology for domain-independent semantic role labeling using language independent techniques. This project resulted in very significant performance improvements in a Semantic Role Labeling system for both English and Chinese, yielding state-of-the-art results (Pradhan et. al. 2004a, Pradhan et. al. 2004c, Pradhan et. al. 2005a Pradhan et. al. 2005b, Hacıoglu 2004). The semantic role labeling system was extended to be able to process nominal predicates as well as verb predicates (Pradhan et. al. 2004b, Xue 2006). We demonstrated that using semantic role labels improved performance over baseline results on a simple question answering system. Applications to Machine Translation were also investigated (Gildea et. al. 2006, Liu and Gildea 2006, Zhang and Gildea 2006a, Zhang and Gildea 2006b, Zhang et. al. 2006, Ding and Palmer 2005, 2006). Thus far 27 publications have resulted from this effort.

Ron Cole has been PI of prior NSF Challenge and ITR grants, and co-PI of an NSF IERI grant that supported educational projects. The NSF Challenge grant supported development of the CSLU Toolkit, which was used to develop vocabulary tutors that dramatically improved the speech and language skills of children with hearing loss. The results were featured on ABC TV's Prime Time and the NSF Home page. The CSLU Toolkit is freely distributed to researchers and has been installed in over 30,000 sites in 130 countries and is widely used in research and education. The toolkit was licensed to Animated Speech Characters Inc and is the basis for the company's learning products for students with cognitive disabilities. The NSF ITR grant and IERI grants supported development of Foundations to Literacy (FtL), a reading program that improves word recognition and comprehension skills through conversational interaction with a lifelike computer character. FtL has been licensed for commercialization to a startup company. The tools and technologies developed under the ITR and IERI grants have led to grants from IES and NIH that have led to virtual therapy programs for individuals with aphasia and Parkinson disease.

PERSONNEL

Principal Investigator Wayne Ward is Director of CSLR and co-founder of Boulder Language technologies. He has been PI on two DARPA projects that produced advanced dialog systems. He developed the Phoenix system for robust semantic information extraction from Spoken Dialogues and text. Since 2001, he has been working in the area of automatic statistical shallow semantic annotation for Question Answering. **Co-PI Ron Cole** is Director and Research Professor at Boulder Language Technologies, which he co-founded with Wayne Ward in February 2007. He established the Center for Spoken Language Understanding at Oregon Health & Science University (previously OGI) and co-founded the Center for Spoken Language Research at CU. Dr. Cole has managed large scale development projects under NSF and NIH support for the past 10 years. He recently organized two NSF-sponsored workshops on Virtual Human Systems. **Co-PI Sarel van Vuuren**, is a Research Associate CSLR. He has led software development of virtual human tutoring and therapy systems that have undergone successful clinical trials with individuals with dyslexia, aphasia and Parkinson disease. **Margaret G. McKeown (Investigator)** is a Senior Scientist at LRDC at the University of Pittsburgh. Her research focus has been the study of students' comprehension from school texts, and co-developed QtA. Dr. McKeown will devote 25% of her time to the project training project staff and analyzing human and computer tutoring sessions. **Larry Malone** is co-director of FOSS and ASK projects at the Lawrence Hall of Science (LHS) at UC Berkeley. He has worked in STEM projects for 40 years. Larry will work with the project team to design and align My Science Tutor to FOSS science investigations in terms of content, pedagogy and formative and summative assessment.

The consulting team (**biographies attached**) also includes **Eric Hamilton** (Director of the Center for Research on Learning and Teaching at the US Air Force Academy), **Finbarr Sloane** (Assoc. Professor of Education at Arizona State), **Linda DeLucchi** (co-director of FOSS and ASK projects) and **Samantha Messier**, PhD, (educational researcher and Boulder Valley School District Science Coordinator). These individuals round out the expertise in assessment, software development, educational research methods, intervention scaling, and science education that are crucial to the success of this effort.

REFERENCES

Allington, R. L. (2001) What really matters for struggling readers: Designing research-based programs. New York: Longman.

America Reads (1997) Available Online:

<http://www.ed.gov/inits/americanreads/resourcekit/miscdocs/tutorwork.html>

Atkinson, R. K. (2002), "Optimizing Learning from Examples Using Animated Pedagogical Agents," *Journal of Educational Psychology*, 94, 416-427.

Baylor, A. L. & Ryu, J. (2003). Does the presence of image and animation enhance pedagogical agent persona? *Journal of Educational Computing Research*, 28(4), 373-395.

- Baylor, A. L. & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education*, 15(1).
- Beck, I. L., McKeown, M. G., Worthy, J., Sandora, C. A., & Kucan, L. (1996) "Questioning the author: A year-long classroom implementation to engage students with text", in *The Elementary School Journal*, 96(4), 387-416.
- Beck, I., and McKeown, M. (2006). Improving comprehension with Questioning the Author: A Fresh and Expanded View of a Powerful Approach. Scholastic.
- Black, P. and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–73.
- Bloom, B.S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring, *Educational Researcher* 13, pp. 4-16.
- Bransford, John D., Brown, Ann L. and Cocking, Rodney R. (Eds.) (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.
- Bransford, J.D., Vye, N., Kinzer, C., Risko, V. (1990). Teaching thinking and content knowledge: Toward an integrated approach. In B. Jones & L. Idol (Eds.). *Dimensions of thinking and cognitive instruction* (pp. 381-413). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Bransford, J. D. (1993). Who ya gonna call? Thoughts about teaching problem-solving. In P. Hallinger, K. Leithwood, & J. Murphy (Eds.), *Cognitive perspectives on educational leadership* (pp. 171-191). New York: Teachers College Press.
- Bybee, R.W. (1996) The contemporary reform of science education. In: J. Rhoten & P. Bowers; (Eds.) *Issues in Science Education* (Arlington VA) 1-14.
- Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25, 471–533.
- Cohen, P.A., Kulik, J.A., & Kulik, C.L.C. (1982) "Educational outcomes of tutoring: A meta-analysis of findings", *American Educational Research Journal*, 19, 237-248.
- Cole, R., Wise, B., Van Vuuren., S. (2007a). How Marni teaches children to read. *Educational Technology*.
- Cole, R., Angela Halpern A., Ramig, L. van Vuuren, S., Ngampatipatpong, N & Jie Yan. A Virtual Speech Therapist for Individuals with Parkinson Disease. *Educational Technology*. *Educational Technology*.
- Cromley, J.G., & Azevedo, R., What Do Reading Tutors Do? A Naturalistic Study of More and Less Experienced Tutors in Reading DISCOURSE PROCESSES, 40(2), 83–113.
- CSLRSYS (2006). CSLR systems described online: http://cslr.colorado.edu/beginweb/vt_th/vt_th.html
- EdWeb (2005). Department of Education MATHEMATICS AND SCIENCE EDUCATION RESEARCH GRANTS PROGRAM, CDA NUMBER: 84.305. <http://www.ed.gov/about/offices/list/ies/programs.html>.

Ferguson, R. F. (1991) "Paying for public education: New evidence on how and why money matters", in Harvard Journal on Legislation, 28, 465-498.

FossInfo (2006). The FOSS Web sites (<http://www.fossweb.com>; <http://lawrencehallofscience.org/FOSS/>) provide a wealth of information and resources for parents, educators, researchers and other interested parties. The FOSS K-8 matrix with summaries of all the modules and courses can be found at <http://lhsfoss.org/scope/index.html>. The most recent edition of FOSS was developed for the 2006 California Science Adoption and the summaries of the program can be found at <http://www.fossweb.com/CA/>. Delta Education is the publishing partner that works with the FOSS research team to provide professional development for new implementers.

FTL Surveys (2006). Teacher and Student Histograms available online: http://cslr.colorado.edu/beginweb/documents/ftl_student_teacher_surveys.html

Gee, J.P. (2004) *Situated Language and Learning: a critique of traditional schooling*. London: Routledge.

Graesser, A.C., Hu, X., Susarla, S., Harter, D., Person, N.K., Louwerse, M., Olde, B., & the Tutoring Research Group. (2001) "AutoTutor: An Intelligent Tutor and Conversational Tutoring Scaffold", in Proceedings for the 10th International Conference of Artificial Intelligence in Education San Antonio, TX, 47-49.

Insights (2007) Developer: Educational Development Corporation, Newton, MA. Publisher: Kendall/Hunt, <http://www.kendallhunt.com/index.cfm?TKN=5C7C9D84-306E-01A4-A2C3D4835490C611&PID=219&PGI=137>

Johnson, W., Rickel, J., & Lester, J, (2000). Animated pedagogical agents: Face to face interaction in interactive learning environments. *International Journal of Artificial intelligence in education*, 11, 47-78.

King, A. (1991). Effects of training in strategic questioning on children's problem-solving performance. *Journal of Educational Psychology*, 83, 307-317.

Klentschy, M. (2007) Improving Student Achievement in K-12 Science: Linking District Decisions to State Assessments. Proceedings of NSF-Sponsored Workshop "Science and English Language Learners." St. Louis, March 31, 2007.

Klentschy, M. (2002) Helping English Language Learners Increase Achievement Through Inquiry-Based Science Instruction. *J Bilingual Research Journal* vol. 26. http://www.fossworks.com/pdfs/Helping_ELL.pdf.

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

Kintsch, E. (2005). Comprehension theory as a guide for the design of thoughtful questions. *Topics in Language Disorders*, 25 (1), 51-64.

- Madden, N.A., & Slavin, R.E. (1989) "Effective pullout programs for students at risk", in *Effective Programs for Students At Risk*, R.E. Slavin, N. L. Karweit, and N.A. Madden, eds. Boston: Allyn and Bacon.
- Malone, L, Long, K. De Lucchi, L. (2004). "All Things in Moderation." *Science and Children*, February 2004.
- Malone, L., Long, K. (2006). "Assessing Science Knowledge (The ASK Project)." FOSS Newsletter #27, University of California, Berkeley, Spring 2006. Available online: <http://lhsfoss.org/newsletters/last/FOSS27.assessing.html>
- Mayer, R. (2001) *Multimedia Learning*. Cambridge, UK: Cambridge University Press.
- McKeown, M.G., & Beck, I.L. (1999). Getting the discussion started. *Educational Leadership*, 57 (3), 25-28.
- McKeown, M. G., Beck, I. L., Hamilton, R., & Kucan, L. (1999). "Questioning the Author" Accessibles: Easy access resources for classroom challenges. Bothell, WA: The Wright Group.
- Moreno, R., Mayer, R.E., Spires, H.A., Lester, J.C., (2001). The Case for Social Agency in Computer-Based Teaching: Do Students Learn More Deeply When They Interact With Animated Pedagogical Agents? *Cognition and Instruction*, 19(2), 177–213.
- 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*. Symposium presented at the annual meeting of the American Educational Research Association, Montreal, Canada.
- NAEP, (2002). <http://nces.ed.gov/nationsreportcard/>
- Nass C. & Brave S. (2005). *Wired for Speech: How Voice Activates and Advances the Human- Computer Relationship*. MIT Press, Cambridge, MA.
- National Academies, (2006). *Taking Science to School: Learning and Teaching Science in Grades K-8*. Committee on Science Learning, Kindergarten through Eighth Grade, Richard A. Duschl, Heidi A. Schweingruber, and Andrew W. Shouse, Editors. <http://www.nap.edu/catalog/11625.html>
- National Research Council (1999). *How people learn: Brain, mind, experience, and school*. Committee on Developments in the Science of Learning. J.D. Bransford, A.L. Brown, and R.R. Cocking (Eds.). Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. J.W. Pellegrino, N. Chudowsky, and R. Glaser (Eds.), Committee on the Foundations of Assessment. Washington, DC: National Academy Press.

- NSRC (1996). National Science Resources Center. Resources for Teaching Elementary School Science. National Academy Press: Washington, DC.
- NSRC (1997). National Science Resources Center. Science for All Children: A Guide to Improving Elementary Science Education in Your School District. National Academy Press: Washington, DC.
- NSTA (1996). National Science Teachers Association. Pathways to the Science Standards: Guidelines for Moving the Vision into Practice, Elementary School Edition (Ed. Lawrence F. Lowery). NSTA: Arlington, VA.
- Paris, S. G., Wasick, B. A., & Turner, J. C. (1991). The development of strategic readers. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of Reading Research: Volume II* (pp. 609-640). New York: Longman.
- Pellom, B., Ward, W., and Pradhan, S. (2000) "The CU Communicator: An architecture for dialogue systems," in International Conference on Spoken Language Processing (ICSLP), Beijing, China.
- Pellom, B. (2001) "SONIC: The University of Colorado Continuous Speech Recognizer", University of Colorado, tech report #TR-CSLR-2001-01, Boulder, Colorado, March.
- Pellom, B., Hacıoglu, K. (2003) "Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, April.
- Pressley, M and McCormick, C B, (1995). Advanced educational psychology for educators, researchers and policymakers, HarperCollins, New York.
- Reeves, B., & Nass, C. (1996). *The Media Equation*, NY: Cambridge University Press.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications.
- Ruiz-Primo, M.A., Ayala, C., & Shavelson, R.J. (2004) Evaluating students' science notebooks as an assessment tool. *International Journal of Science Education*, V. 26, No. 12, 1477-1506.
- Snow, R. (1989). Aptitude-Treatment Interaction as a framework for research on individual differences in learning. In P. Ackerman, R.J. Sternberg, & R. Glaser (ed.), *Learning and Individual Differences*. New York: W.H. Freeman.
- Soter, A.O., Rudge, L. (2005). What the Discourse Tells Us: Talk and Indicators of High-Level Comprehension, Annual Meeting of the American Educational Research Association, Montreal, Canada, pp. 11-15.
- Spradley, James P. 1980. *Participant observation*. Orlando, FL: Harcourt Brace Jovanovich College Publishers.
- STC (2007) Science and Technology for Children (STC) Developer: National Science

- Sweet, A. P. and Snow, C. E. (Eds.). (2003) Rethinking reading comprehension. New York: Guilford Press.
- Topping, K., & Whitley, M. (1990) "Participant evaluation of parent-tutored and peer-tutored projects in reading", in *Educational Research*, 32(1), 14-32.
- Valdez, J.D. (2001) The Effect of Inquiry-based Science Teaching on Standardized Reading Scores. <http://sustainability2002.terc.edu/invoke.cfm/page/143>.
- VanLehn, K. & Graesser, A. C. (2001). Why2 Report: Evaluation of Why/Atlas, Why/AutoTutor, and accomplished human tutors on learning gains for qualitative physics problems and explanations. Unpublished report prepared by the University of Pittsburgh CIRCLE group and the University of Memphis Tutoring Research Group.
- VanLehn, K., Lynch, C., Taylor, L., Weinstein, A., Shelby, R., Schulze, K., Treacy, D., & Wintersgill, M. (2003). In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Intelligent Tutoring Systems 2002* (pp. 367-376). Berlin, Germany: Springer.
- Van Vuuren, S. (2007). Technologies that power pedagogical agents and visions for the future. *Educational Technology*.
- Vaughn, S., Marie Tejero Hughes, Sally Watson Moody, and Batya Elbaum. (2001). Instructional Grouping for Reading for Students with LD: Implications for Practice. *Intervention in School and Clinic*, January 2001, Vol 36, No. 3. (pp.131-137) <http://www.readingrockets.org/article/203>
- VH-Web (2007) http://cslr.colorado.edu/beginweb/vt_th/vt_th.html.
- Ward, W., (1994) "Extracting Information From Spontaneous Speech", In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Sept. 1994.
- Ward, W., Pellom, B. (1999). The CU Communicator system. Proceedings of the 1999 IEEE Workshop on Automatic Speech Recognition and Understanding Workshop (ASRU), Keystone, Colorado, 1999.
- Weston, T. J. (2004). Formative evaluation for implementation: Evaluating educational technology applications and lessons. *American Journal of Evaluation*, 25(1), 51-64.
- Wise, B.; Cole, R.; van Vuuren, S.; Schwartz, S.; Snyder, L.; Ngampatipatpong, N.; Tuantranont, J.; & Pellom, B. (in press). Learning to Read with a Virtual Tutor: Foundational exercises and interactive books. In Kinzer, C. & Verhoeven, L. (Eds). *Interactive Literacy Education*. Mahwah, NJ: Lawrence Erlbaum. Available: http://cslr.colorado.edu/beginweb/virtual_tutor/virtual_tutor.pdf