

Improving Science Learning through Tutorial Dialogs

Submitted to IES Cognition and Learning

November, 16, 2006

Ron Cole

Barbara Wise

Wayne Ward

Sarel van Vuuren

Center for Spoken Language Research, University of Colorado

Margaret G. McKeown

Learning Research and Development Center, University of Pittsburgh

Finbarr Sloane

Arizona State University

Larry Malone

Linda DeLucchi

Kathy Long

Lawrence Hall of Science, University of California at Berkeley

Samantha Messier

Boulder Valley School District

Improving Science Learning through Tutorial Dialogs

The RFA goal of the proposed work is development. Its purpose is to evaluate the integration of tutorial dialogs into elementary science curricula and to compare their use in large classroom vs. small group settings. The purpose is also to evaluate the use of a virtual tutor that emulates an effective human tutor in conducting the dialogs. Work by this team has already established that students engage with and learn from virtual tutors in reading programs (Cole et al, 2006). The current project aims to demonstrate the potential of tutorial dialogs to improve children's science learning in schools that have previously performed poorly with FOSS (Full Option Science System), a well-structured science program that is effective with many but not all students. The system will use an effective approach to dialog interaction to stimulate students to learn and integrate new concepts with prior knowledge in order to deepen and expand the knowledge that was presented in class. That system, called Questioning the Author (QtA, Beck & McKeown, 2004), uses systematic dialog interaction to facilitate deep learning; it has already demonstrated effectiveness in improving reading comprehension and is in widespread use in U.S. classrooms. The virtual tutoring system will closely resemble the tutorial dialogs produced by human tutors trained in the Questioning the Author method. Initially, we will develop and refine the system with the help of teachers, the FOSS developers, QtA experts, and the help of 3rd, 4th and 5th grade students from the Boulder Valley School District with varying ethnic, racial, and economic diversities. We will compare student experiences and learning gains on six science subjects for students who interact with both human and virtual tutors in standard classroom or small group settings. We will use observational and interview measures about usability and likeability as well as pre-post assessment measures within correlational and quasi-experimental designs. In the final year of the project, we will assess the feasibility and potential of three dialog treatments by comparing the learning of science modules in an experimental design with up to 672 3rd, 4th and 5th grade students. The students will reside in 8 classrooms in each of three grade levels in the Boulder Valley School District in Colorado which has a mixed racial/ethnic composition. We will analyze their performance on content-related pre-, post-, and delayed post assessments for generalization of learning. Students will be randomly assigned to receive either (a) standard classroom instruction and support, (b) classroom instruction with support that incorporates QtA dialogs in a large group, (c) small-group tutoring with QtA with a well-trained human tutor, or d) small group interaction with the computer-based QtA tutoring system. We hypothesize that the computer-based QtA intervention and the small group QtA will both produce learning gains significantly greater than large group support with QtA in classrooms, and that all dialog treatment conditions will produce greater learning gain than the untrained classroom control. Scores from pre and post measures and follow-up measures will be used as dependent variables for analysis. We will use statistical techniques that leverage our capacity to randomly assign. 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. We expect gains from small group QtA instruction with the computer system to be intermediate between large group QtA with well-trained tutors and the small groups with human QtA tutors, (and much less expensive for schools to provide). In addition, we will continue to analyze interview and questionnaire data on usability and likeability. Successful outcomes will demonstrate the potential of a scalable and cost-effective tool for improving science learning in a program used by over 2 million children today.

RESEARCH NARRATIVE

1. Significance

Overview

The primary goal of the proposed work is to demonstrate that tutorial dialogs integrated into science instruction in elementary schools can produce significant and meaningful learning gains for third, fourth and fifth grade students. We propose to achieve this goal by demonstrating learning gains in six areas of science, two each in 3rd, 4th and 5th grades in the Boulder Valley School District in Colorado, representing half of the science curricula at each grade level in one school year. These dialogs will be designed to assess what students do and don't know about the science they are learning during classroom science investigations, and to guide them to think and reason about key facts and concepts and the connections between them to arrive at accurate mental models that support transfer of knowledge to new tasks. The process of dialog design will be informed by a significant body of prior research that provides insights about the main challenges that students have in mastering science concepts in each science domain and by proven 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). The result of the dialog design process will be a set of dialogs that are designed to assess and improve individual student's comprehension of science concepts following classrooms science investigations using the FOSS science program. We will investigate and compare the potential benefits of tutorial dialogs conducted immediately after students complete well-designed science investigations in the classroom by comparing three dialog-based interventions to a classroom control condition: tutorial dialogs in classroom settings with a QtA expert, tutorial dialogs in small group settings with a QtA expert, and tutorial dialogs in small group settings with a virtual tutor, a lifelike computer character that has been programmed to closely emulate the dialog interactions of the human experts. The research is thus designed both to assess the feasibility and potential of the proposed computer-based intervention, to demonstrate the potential of a powerful approach to learning through dialog interaction in classrooms and small groups to improve science learning, and to advance scientific knowledge by providing new insights about how students learn through dialog interaction in different social contexts—in classrooms and in small groups with both human and virtual tutors.

Two main assumptions that drive our research are (1) learning benefits greatly from individualized instruction, and (2) learning benefits greatly from instructional approaches that use well designed tutorial dialogs that stimulate critical thinking and relate new information to prior knowledge. Research reviewed below shows that one-on-one tutoring provides significant learning gains relative to classroom instruction over a wide range of tutoring approaches and student populations (e.g., Bloom, 1984; Cohen et. al., 1982) and that instructional methods that incorporate systematic, well designed dialogs produce significant learning gains (Cohen et al., 1982; King, 1991; Beck & McKeown, 2002; Murphy & Edwards, 2005). Well designed tutoring dialogs can guide students to a deep understanding of new information presented in a text, lecture, movie or multimedia presentation by helping them make connections between the main ideas in the presented material and their own knowledge. Instructional techniques that use spoken dialogs to help students make connections between new ideas and their knowledge of the world are consistent with theories of comprehension that describe learning as an active process

that involves constructing new mental models by integrating concepts presented in a discourse with prior knowledge (Kintsch, 1988, 1998; Bransford, et al., 1999; Snow, 2002).

Importance of the Problem

In the 2002 National Assessment of Educational Progress (NAEP), only two percent of U.S. students attained advanced levels of mathematics or science achievement by Grade 12” (EdWeb, 2005). A significant factor contributing to this national crisis is the inability of students to comprehend and learn from science texts, where much of the information resides that students are required to learn. While many students may appear to learn to read and understand text by third grade, evidence shows that their apparent competence is often an illusion—as texts become more challenging in fourth grade, many students cannot read nor understand them (Meichenbaum & Biemiller, 1998; Sweet & Snow, 2003).

A number of studies have shown that most teachers do not engage students in effective learning dialogs in classrooms. Students ask about 3 questions per hour in classrooms. While teachers ask questions more frequently, between 30 to 120 questions per hour, with a mean of 69 per hour (Graesser and Person, 1994), they do not typically ask questions that stimulate deep thinking or that lead to meaningful teacher-student dialogs (see Graesser and Person, 1994 for review). The Questioning the Author approach that will be used in the proposed research was designed precisely to address this problem by training teachers to engage students in dialogs that challenge them to think and reflect about what they are learning and how it relates to their knowledge of the world. Recent research has shown that the use of QtA dialogs can change the process by which students learn and lead to significant learning gains in different subject areas.

Effective intelligent tutoring systems exist today in the laboratory that use natural, text-based dialog interaction with a Virtual Tutor. These systems produce up to one sigma (one grade level) learning gains with high school and college subjects relative to normal classroom instruction Graesser et al., 2001; VanLehn & Graesser, 2001; VanLehn et al., 2002). Recent research with college students using intelligent tutoring systems has also shown advantages of Socratic dialogs over alternative approaches (Rose et al., 2001). There are however, no systems yet developed that teach cognitive strategies to children for reasoning about and making sense of new ideas through interaction with a virtual tutor. The work proposed here, if successful, will result in an accessible, inexpensive, scalable and effective tutoring program that can provide one critical component of an overall solution to improved education and learning. From a cost-benefit perspective, computers and associated learning software can provide a relatively inexpensive solution in today’s education system. Successful outcomes of the proposed work could therefore have significant theoretical, empirical and practical benefits.

Specific Benefits of the Proposed Work to Science Education in the United States

A major theme and expected benefit of the proposed work is to develop treatments that help individual students to overcome barriers to science learning. While we expect each of the proposed tutoring treatments to benefit the vast majority of students, our work also addresses the question: *how can we improve science learning for low performing students who fail to achieve their potential* even when they are immersed in an exceptional science program (FOSS) that has been demonstrated to produce significant overall gains in student achievement within their school district?

FOSS (Full Option Science System) is based on the idea that “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.” This philosophy is consistent with a large body of theory and research, and with recommendations presented in reports sponsored by the nation’s top science and educational organizations (e.g., NRC, 1999; 2001; National Academies, 2006).

FOSS is in use 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. FOSS was the first non-textbook curriculum to make the California adoption list in 1992 and has just recently 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 100 largest U.S. school districts. FOSS is cited as an exemplary program in publications by nationally recognized organizations in the science reform movement: National Science Resources Center - Resources for Teaching Elementary School Science (NSRC, 1996), and Science for All Children (NSRC, 1997); and the National Science Teachers Association - Pathways to the Science Standards (NSTA, 1996).

FOSS has been under development since 1988 at the Lawrence Hall of Science, University of California at 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 to the National Science Education Standards, and to science standards in several states. Pointers to additional information about the FOSS program are provided in the references under FossInfo (2006).

There is no question that FOSS is a well designed and highly effective program. 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. 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, 2006 Colorado CSAP science scores of 1, 954 fifth graders 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), but 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.

FOSS researchers are working hard to provide teachers and students with more effective formative assessments and other learning tools that can be used to guide instruction and improve learning. However, it is possible that some students, because of cognitive, linguistic or cultural factors (that may affect vocabulary, background knowledge, motivation, self confidence or learning strategies) are unable to benefit sufficiently from classroom participation in well-designed science investigations. The authors of this proposal believe that small-group or individualized instruction using tutorial dialogs that incorporate principles of multimedia learning can benefit many of these students. If this outcome is achieved in the proposed work,

the integration of tutorial dialogs into the FOSS learning program in the future could benefit millions of students and improve student achievement in the thousands of classrooms currently using FOSS in the U.S.

Theoretical Framework and Empirical Rationale for Tutorial Dialogs

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 model provides the theoretical 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 in the context of the present project. 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 the 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 that students form (superficial understanding versus an adequate situation model) is determined by the comprehension strategies students use (Paris et al., 1991; Pressley & McCormick, 1995).

Within this theoretical framework, our work focuses on the role of dialog interaction in science learning. 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. 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.

Benefits of Tutorial Instruction

Accommodating different learners’ needs through quality tailored instruction via mixed-initiative dialog interaction is the core of the proposed project. The focal aspect of teaching and learning on which we propose to base our program is *the instructional conversation*. In 1984, Benjamin Bloom determined that the difference between the amount and quality of learning for students who received classroom instruction and those who received either one-on-one or small group tutoring was 2 standard deviations. These results were replicated by Bloom’s students across different content areas and subject populations, including middle school students.

In the two decades since Benjamin Bloom reported a two sigma advantage of one-on-one tutoring over classroom instruction across different student populations and subject areas,

evidence that tutoring works has been obtained from dozens of well designed research 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) have provided insights about factors that improve tutoring outcomes, including frequent, regular and well structured tutoring sessions, coordination with classroom activities, well trained tutors, and dialogs that are structured to stimulate thinking and knowledge construction (America Reads, 1997).

Benefits of Questioning the Author

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. The approach uses open-ended questions to initiate discussion (What is the author trying to say?) to help students focus on the author's message (That's what she says, but what does she mean?) to help students link information (How does that fit with what the author already told us?) and to help the teacher guide the student toward comprehension of the text. The approach is guided by a planning process that involves identification of major conclusions by the teacher and that helps the teacher move the dialog toward learning goals while encouraging creative thinking.

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.

In a recent IES funded study conducted by Ian Wilkinson and his colleagues that 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).

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, children learned more and reported more satisfaction using programs that incorporated virtual humans (e.g., Moreno, 2001, Atkinson, 2002; Baylor et al., 2003, 2005). Lester terms the tendency to treat pedagogical agents like real teachers “The Persona Effect” (Lester et al, 1997). He argues that the social bond that students establish with animated agents seems to motivate students to please the virtual teacher and work harder on learning tasks. Our own research, based on surveys of over 400 students conducted during the past two school years, supports this result. Over 90% of students surveyed believe that Marni (the virtual teacher) acts like a real teacher and that Marni teaches them to read. Over 80% report that they trust Marni and believe that she cares about them. When K-3 students who used the program during the 2005-2006 school year were asked: “Do you think that answering the questions Marni asked you about each story helped you to learn?” 222 students answered “yes” and 17 answered “no.” 98% of students responded “yes” to the question “Do you wish you could have had more time working with Marni?”

Benefits of Multimedia Presentations

The results of research conducted by Richard Mayer and his colleagues at UC Santa Barbara provide clear guidelines for designing multimedia presentations to optimize learning of science concepts (see Mayer, 2001 for a review). One of the most important results to emerge from Mayer’s work is that people learn scientific concepts best when they are able to view a series of illustrations or an animation that is accompanied by a simultaneous spoken explanation of the concepts presented visually. Mayer’s work, replicated and extended by a number of independent researchers, shows that narrations of illustrations or animations results in improved retention and transfer of concepts to new problems. Based on this work, we plan to investigate benefits of incorporating narrations of visual materials at appropriate points in tutorial dialogs to help students understand science concepts that they have difficulty explaining adequately.

2. Research Narrative

2.1 Prior Work on Spoken Dialogs and Virtual Tutors at CSLR

Development of the human and virtual tutoring systems described below builds on over two decades of research by Ron Cole and Wayne Ward and their colleagues in two major areas-- *natural spoken dialog systems* (Wayne Ward) that support unconstrained, mixed-initiative dialogs between people and computer systems, and *intelligent tutoring systems* that teach reading and comprehension through conversational interaction with a virtual tutor (Ron Cole).

Development of these systems is directly related to the work proposed below for both human and virtual tutoring treatments, as the computer systems we develop rely heavily on methodologies designed to understand the nature of human-human communication. To date, CSLR has developed about a dozen systems, described on the CSLR Web site (CSLRSYS, 2006), that use virtual humans to provide tutoring, training or therapy; in each case, development of these systems was based on participatory design methodologies, in which both domain experts (reading researchers, teachers, therapists) and end users (students, patients) participated in the design and testing of the systems. These methodologies will be incorporated into the design of both the human and virtual human treatments described below. In the proposed work, the participatory design team will consist mainly of science educators, educational researchers, system developers, and students.

On the technology side, the proposed intervention builds on state of the art tools, technologies and system architectures developed at CSLR in areas of speech, language and character animation technologies and a set of established methodologies for developing and testing tutoring systems to assure ease of use and user satisfaction. The performance of systems developed and tested at CSLR, combined with the results of an initial pilot study described below, suggests that the proposed computer intervention can be developed with high certainty. In the remainder of this section, we describe the tools, technologies and systems that we will use to support natural spoken dialogs with virtual tutors. We note that all of these tools are being made available to the research community free of charge as they are ready for distribution.

Interactive Books: Software Toolkit and Platform for Spoken Dialogs with Marni

During the past five years, we have developed an engaging and effective learning tool, *Foundations to Literacy (FtL)*, which teaches reading and comprehension to K-3 students through conversational interaction with Marni, a virtual tutor. The program has been used by over 2000 students in 50 classrooms in 5 school districts in Colorado elementary schools. The system is highly stable; support for over 120 computers in this project is provided by one full time staff member. The program has received positive reviews from teachers and students. Summative evaluation revealed significant learning gains in word recognition and comprehension across the three years of the project, with moderate effect sizes following five to ten hours on the program (Cole et. al., In press; Cole et al., 2003, Wise et al., in press, FTL Surveys, 2006).

Foundations to Literacy consists of three integrated components: (a) a Managed Learning Environment that monitors, logs and stores all student-system interactions, displays student progress within the program and manages an individual study plan for each student; (b) Foundational Skills Reading Exercises, which teach and practice basic reading skills that



Figure 1. Interactive Book in multiple choice question and reading modes.

underlie reading; and (c) Interactive Books, which support activities designed to teach students to read accurately, fluently and with good comprehension.

Interactive Books (see Figure 1) enable developers to author a wide range of user interactions with the Virtual Tutor, including: (a) having the story, or any portion of it, narrated by virtual tutor, which produces accurate visual speech synchronized with recorded speech, head movements and facial expressions, (b) enabling the student to click on individual words or sentences while reading to have them spoken by the Virtual Tutor, (c) providing feedback to the student while they read out loud by tracking their reading position using automatic speech recognition technology (Hagen et al., 2003, 2004) and (d) having the student respond to questions posed by the Virtual Tutor (by clicking on objects in images, answering multiple choice questions, or saying or typing responses).

Interactive Books provide a powerful and flexible software environment for authoring and conducting tutorial dialogs in a rich multimedia environment. The tools allow Marni to be placed anywhere on the screen (at any size) to converse with students. A text markup language developed at CSLR enables designers to control Marni's emotions, head movements and facial expressions at any time in a dialog while she is speaking or listening. The authoring tools also enable designers to present sounds, music, graphics or flash animations at any point in a dialog. In addition to speaking and listening, Marni can focus her attention on any part of the computer screen so she can, for example, look at an illustration while explaining a science concept or watch an animation while narrating it. Thus, when the dialog system determines that the student is struggling and the conversation needs to be grounded, Marni might say "Let's look at the water cycle again and then we will continue to discuss it." Examples of narrated flash animations used to present science concepts (metamorphosis, movement of land masses, water cycle, formation of volcanoes, etc.) can be viewed on the CSLR Web site (Flying-Volando, 2006).

The integration of spoken dialog systems that incorporate speech recognition, natural language understanding, language generation and dialog modeling into Interactive Books to support spoken dialog interaction with virtual tutors in rich multimedia environments is already underway at CSLR under support from an NSF Computer Research Infrastructure grant. The goal of the work conducted under this grant is to provide scientists with a free and powerful tool for conducting research that supports spoken dialogs with virtual humans.

In the current implementation of *FtL*, Interactive Books for grades 1-5 assess and train comprehension through *thinking multiple choice (MC) questions*, developed by our research team (Kintsch E, 2005). The design of thinking MC questions was motivated by the Discourse

Comprehension theory and informed by research conducted by McKeown and Beck (1996) and others. 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. Thinking MC questions in Interactive Books include gist questions that target the main ideas and most important events in the story, and inference questions that target information not explicitly stated in the text. These inference questions require the learner to either (a) connect ideas from sentences in different parts of the story (textbase knowledge), or (b) stretch beyond the story to draw a conclusion or implication beyond what the author has explicitly stated. In the proposed work, tutorial dialogs will be used to assess and train comprehension. However, thinking multiple choice questions may be used to assess students' knowledge of specific content quickly and accurately before or after tutorial dialogs.

CAT: Tools and Technologies Enabling Tutorial Dialogs

Development of the spoken dialog systems that form the basis of tutorial dialogs builds on tools, technologies and system architectures already in place and well tested at CSLR. The initial baseline system will build on the technology servers used in the CU Communicator project (Ward & Pellom, 1999; Pellom, et al., 2000). The tools from that system have been used to create CAT, the CU Conversational Agent Toolkit (Cole et al., 2003). CAT provides a general-purpose platform, a set of technology modules, and tools for researching and developing advanced dialog systems—systems that enable completely natural and unconstrained mixed-initiative conversational interaction with users in specific task domains. In mixed-initiative dialogs (as distinguished from finite state grammars), 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 gracefully 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.

Speech Recognition: Our large vocabulary continuous speech recognition system, *Sonic* (Pellom, 2001; Pellom and Hacioglu, 2003), developed at CU, produces accurate recognition of children and adult speech, and can be quickly trained and optimized for new dialogs. In addition to large vocabulary speech recognition, the recognizer supports both keyword/phrase spotting and constrained grammar-based speech recognition. It provides an integrated environment that incorporates voice activity detection (VAD) and speech enhancement as well as various feature and model-based speaker adaptation and normalization methods. The recognition architecture provides support for rapid portability to new languages. *Sonic* has been ported from English to the French, German, Italian, Japanese, Spanish, and Turkish languages. It has also been trained on children's speech for use in interactive books.

Information Extraction: We use the Phoenix parser (Ward, 1994) to map the speech recognizer outputs onto a sequence of semantic frames. These extracted frames represent the meaning of the utterances. Since its development in 1990, Phoenix has been used for many applications in limited domain spoken language understanding. It has proven to be robust and

easy to author. It is able to parse spoken input that includes false starts, filled pauses, hesitations, repeated words, corrections and ungrammatical constructions. The Phoenix parser was used in the CU Communicator system, which was one of the top performing systems in DARPA competitions comparing systems fielded by top U.S. labs. The system, which is similar in complexity to the tutorial dialogs we propose to develop, enabled callers to create air travel itineraries that included lodging and rental cars. Our system achieved 95% task completion rates and good user satisfaction ratings (Pellom et al. 2000).

Dialog Control: The Dialog Manager (DM) controls the system's interaction with the user. It is responsible for several different functions: (a) Maintaining a semantic context for a session; (b) Receiving extracted information from the current input and integrating new information into the context. This includes verification based on confidence assessment; ellipsis and anaphora resolution; clarification; and context update; (c) Natural language generation, which includes prompting for information, outputting information to the user, and clarification.

Text-to-Speech Synthesis: The TTS synthesizer receives word strings from the natural language generator and synthesizes them into audio waveforms that can be played back to the user. Our current speech synthesizer servers make use of the general-purpose Festival speech synthesis system (Taylor et al., 1998), as well as a domain-specific variable-unit concatenative synthesizer currently used for the CU Communicator travel-planning system. For the concatenative synthesizer, we use a voice talent to record prompts, sentences, phrases and words specific to the domain (lesson content). The system automatically selects the optimal set of recorded speech from its inventory and splices the pieces together, smoothing them for smooth transitions. In those cases where the pieces are phrases or larger, the speech sounds very natural. We expect that to be the case in this application. Almost all of the output will be pre-recorded phrases or sentence and will sound very natural.

Our basic data structures for representing domain information are frames. Information from the current input is extracted into frames and these newly extracted frames are merged with the current context to create a new context. The CAT Dialog Manager has a scripting capability that allows a developer to control the sequencing of actions. A developer creates the dialog control by creating frames and frame elements (slots in the frames) that are the concepts related to the frame. A semantic grammar is associated with each frame element and represents how users (students in this case would talk about the concept). Templates are associated with each concept that specify the various ways that the system would ask or answer questions about the concept.

The type of processing Phoenix uses to extract information from user input is generally referred to as shallow semantic parsing. Shallow semantics refers to making explicit the underlying predicate argument structure which can be characterized as a semantic frame, together with semantic slots and fillers, i.e., the predicate arguments. Phoenix 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.

We note that several intelligent tutoring systems use Latent Semantic Analysis (LSA) to interpret students' explanations or summaries. The advantage of these systems is their coverage

(or recall) with no specific adaptation to the lesson content. They do have a number of weaknesses, especially: 1) They work better on longer amounts of input and have problems with relatively short answers, and 2) They lack precision. LSA is a bag-of-words based technique. While it is very robust in terms of coverage, it is not precise. LSA assumes that if you are generally using relevant words that you are correct. It is not able to represent the relationships between the entities and concepts, i.e. it can't represent the difference between *the Earth revolves around the Sun* and *the Sun revolves around the Earth*. Shallow semantic parses, such as those produced by Phoenix do represent this difference. They also produce a representation that allows the system to pinpoint what was wrong or under-specified about the answer, which enables us to program appropriate follow-up questions. LSA based systems do not produce such a representation. The responses in the QtA tutorial dialogs will generally be short answers, and much more suited to shallow semantic parsing than LSA.

It is important to consider that the extracted information is not being used to score the student, or give them feedback that they are wrong. It is being used to inform the system as to whether it should discuss the current point further, present information, or move on to another point. Lack of coverage therefore has a less serious effect than if the student were being scored.

2.2 Developing Tutorial Dialogs for Human Tutors

Development of tutorial dialogs conducted by human tutors and virtual tutors is a single integrated process. The first stage of the process is for skilled human tutors to develop effective dialogs for each science concept in a module. Once this is done, transcriptions of approved (effective) dialogs by many different tutors interacting with many different students are used by the system developer to develop, test and refine dialogs conducted by the virtual tutoring system. For ease of exposition, we describe the development process for dialogs conducted by human tutors in this section, and describe the process of developing dialogs for virtual tutors in the next section. In each section we describe the process of developing dialogs first, and then provide an overview of the development plan. Further details of the development process are provided in the budget justification. Tutorial dialog development activities span the first three years of the project, and overlap the first year of assessment. Dialogs used during year-three assessment activities will be transcribed, analyzed and refined for year-four assessment.

Developing effective dialogs by human tutors for a given science concept requires a deep understanding of the science content, the learning objectives, the instructional methods designed to achieve these objectives and the range of learning challenges that students bring to or encounter during the learning task. Once these variables are well understood, it is necessary to design dialogs that 1) engage students in conversations that provide the tutor with the information needed to identify gaps in knowledge, misconceptions and other learning problems and 2) guide students to arrive at correct understandings and accurate explanations of the scientific processes and principles. A related challenge in tutorial dialogs is to decide when students need to be provided with specific information (e.g., a narrated simulation) in order to provide the foundation or context for further productive dialog. Students sometimes lack sufficient knowledge to produce satisfactory explanations, and must therefore be presented with information that provides a supporting or integrating function for learning.

Developing a sufficient number of dialogs to support integrated tutoring sessions following science investigations during about half of the 3rd, 4th and 5th grade school year is an ambitious task. Fortunately, the development process is made tractable by research conducted by the FOSS

development group that identifies the learning objectives and measures the learning process for individual students through formative assessments embedded in the program, by summative assessments developed and tested during the past three years administered before, during and following science investigations in each module, and by the well developed and well tested principles of Questioning the Author that inform, facilitate and constrain the design process.

In the remainder of this section we describe the process of designing, developing and testing tutorial dialogs for use in small-group settings. As noted, the process involves two sequential stages. In the first stage, addressed in this section, dialogs are developed and tested by project staff working with students first in one-on-one and then small group sessions. We note that *one-on-one dialogs are only conducted during the initial phases of the development cycle*; subsequent development occurs in dialogs in small group settings, to reflect eventual treatment condition using the virtual tutor. Tutorial dialog development will be conducted in classrooms (in schools different from those used in the assessment phase of the project) immediately following classroom science investigations. Development of tutorial dialogs during this phase of research will be conducted in close collaboration with Linda De Lucchi and her colleagues at the Lawrence Hall of Science (developers of the FOSS program), and with Margaret McKeown, co-inventor of the Questioning the Author approach. The audio recordings made during these tutoring sessions are recorded, transcribed at the word level, and then scored and annotated by Dr. McKeown's research team to indicate and rationalize excellent dialog interactions and to identify poorly implemented dialogs and explain why they were inadequate and how they should be improved. Development of spoken dialog systems uses these recordings and transcriptions to develop and test initial prototypes.

Dialog design will be influenced strongly by research conducted during the past three years under NSF support by Linda De Lucchi, Larry Malone and Kathy Long at UC Berkeley (Malone et al., 2004; Malone & Long, 2006). The project, called Assessing Science Knowledge (ASK) is based on the premise that 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). 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.

The ASK project is designing two types of assessments, embedded (formative) assessments that are incorporated seamlessly into the science investigations, and benchmark assessments that can be used for formative or summative purposes. "Embedded assessments provide diagnostic information about student learning to both teachers and students as teaching and learning are happening. Embedded assessments generally involved teacher observation (watching students' inquiry practices during investigation activities), looking at written work in science notebooks, and having students engaging in self-reflection."

The work being conducted in the ASK project, which will be completed before the proposed start date of this project, provides a wealth of relevant knowledge that will be used to inform the design of QtA dialogs. For example, the dialogs should, at a minimum, include the constructs that students are expected to learn in each of the three to four parts of each science investigation. As a second example, the dialogs should be designed to anticipate and deal with the specific challenges that individual students encounter in learning concepts associated with the investigations. These challenges are identified in surveys (pretests) administered to each student before each module, and in formative and summative assessments administered during and at the end of each investigation, respectively. In addition to using survey results, formative assessment and summative assessment data will inform the design of dialogs to deal with anticipated learning challenges, we will use these same data to determine which concepts can be presented most effectively through narrated simulations incorporated into the dialog system.

2.3 Developing Spoken Dialogs with a Virtual Tutor

The goal of the virtual tutor 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 to improve science learning in FOSS science investigations.

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 each science investigation in each FOSS module. A question-answer dialog is then designed, based on analysis of QtA dialogs conducted by our project team, to assess the student's knowledge of these concepts.

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 the 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 and queried to refocus student thinking.

In science investigations, the perspective of the "author" in the context of "Questioning the Author" moves from questions about what a specific author (of a text, painting, multimedia presentation) 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 with objects, organisms, and systems, and guiding that sense making to converge on conventional models and understandings, that is, generally accepted (and functional) scientific knowledge. 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...?

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. In comparing the points in the student's answer, the system looks to see if the semantic roles are correct and also if equivalent entities are filling the same roles. Questions are generated about the missing or erroneous concepts to attempt to elicit information about them. For each point, the system has a template associated with the concept that specifies how to generate prompts or questions about the information.

In order to create a tutorial dialog, a developer must:

- Define a set of frames containing the concepts (or points) to be covered by the student's answers. These are referred to as frame elements.
- Specify a template for eliciting information with each frame element.
- Create a context-free grammar for each frame element that specifies the words strings that will be used to match an instance of the element.
- Create a script to interweave presentations with QtA dialogs

The process of developing a tutorial dialog system for each science investigation in a FOSS science module will require an estimated two to three weeks of focused effort by a system developer who has been trained to conduct QtA dialogs. The system developer will work closely with the project's science education researchers (at Berkeley and BVSD) to become knowledgeable about the learning objectives, construct maps and embedded assessments in each science investigation, and learn about the specific challenges that students have in learning the vocabulary and concepts related to each investigation. The system developer will also observe some QtA dialog sessions in person and study videos of tutorial dialog sessions.

With this knowledge base, the developer proceeds to design tutorial dialogs based on the transcriptions of QtA dialogs conducted by our research team with dozens of students. These transcriptions provide the data for designing and testing the spoken dialog system. Dialog design is an iterative design-and-test process; for each sub-dialog that is being developed (corresponding to daily classroom science investigations), the developer designs grammars to parse student responses to questions in the transcribed dialogs. The dialog system is then tested on a new set of transcriptions, and the parser is modified to accurately parse these utterances. In our experience, this process results in acceptable coverage on new dialogs following training on 20 to 30 dialogs. The system is then tested and critiqued by project staff before being field tested with students.

Following the initial phase of development, the system is field tested with students, first in the laboratory, and then in classrooms 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 a staff member (who is expert in QtA dialogs for the science area) who can override the system's dialog responses in order to present the student with a correct response if the system makes an error. By using a "Wizard of Oz," a researcher (behind the "curtain") who can monitor and control the speech of the virtual tutor, we are able to analyze system errors while continuing the dialog with the student. The data collected during these sessions is used to retrain the speech recognizer and improve the coverage of the semantic parser. This process results in a system with acceptable performance. The system logs all transactions (speech files, recognizer output, parser results, system response) during use. Using these data, we are able to transcribe and analyze the performance of individual system components and to analyze and

evaluate the quality of the dialog relative to FOSS construct maps and learning objectives. The speech data and transcriptions are also invaluable resources for training the recognizer's acoustic and language models, for improving coverage of the semantic parser, and for improving the dialog model.

The Wizard of Oz experiments provide an important first-hand mechanism for observing, analyzing and evaluating student interactions with the virtual tutor. The Wizards, who have been trained to be expert QtA human tutors for each science investigation, make judgments about each response that Marni is about to produce during the dialog session. If the Wizard validates the staged response, Marni delivers the programmed response to the student. If the staged response is not validated by the Wizard, Marni 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.

Following the initial Wizard of Oz experiments, conducted in the laboratory over a period of five to six weeks, the resulting program, consisting of tutorial dialogs for each activity within a science investigation, will be tested in classrooms. This dialog testing, observation and recording, transcription, evaluation, refinement and re-testing process will continue in classrooms throughout the first two years of the project. All computer dialogs will be conducted with individual students before being tested with small groups. During this phase of classroom testing, a staff member will observe dialog sessions and intervene as a Wizard to take control of the program if the dialog stalls. The system will then be tested in classrooms with small groups, with students taking turns speaking to Marni, and conferring on answers when needed. The pedagogical dynamics of small group participation will be determined through discussions with the FOSS and QtA experts, and through experiments that investigate different models of interaction, such as having students confer before each response, or enabling students to confer when a student interacting with the virtual tutor requests help.

Results of a 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 Hacıoglu 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.

2. 4 Development Plan & Schedule

As noted above, development of the computer-based dialogs is tightly linked to development, evaluation and validation of dialogs in tutoring sessions with project tutors. For a particular set of dialogs (such as the 4 parts of Investigation 1 of the Magnetism and Electricity module), the process consists of (a) initial design, test and refinement of dialogs by project staff, (b) observation, analysis, evaluation and feedback on dialogs by experts (McKeown, De Lucchi, Malone et al.), (c) transcription of effective dialogs, (d) iterative design and testing of the spoken dialog systems using the speech data and transcriptions, (e) Wizard of Oz experiments, and (f) classroom testing and refinement of dialogs.

During the initial three months of the project, a team of 6 QtA tutors (project tutors) plus the project PI and co-PIs will be trained in QtA dialog techniques by Dr. McKeown in consultation with Dr. Samantha Messier, BVSD Science Coordinator, and FOSS project staff. Training will begin with a three-day workshop. The workshop will start by familiarizing participants (tutor trainees and project PIs) with the theoretical and research background of QtA. The components of QtA discussions will be presented and explored through examples and video clips. Opportunities to analyze transcripts from master QtA teachers will be provided. Participants will take part in QtA discussions and plan a QtA lesson on their own after completing and studying the materials that accompany a FOSS science investigation. Participants will try out their lessons with students. These practice tutoring sessions 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.

Following the introductory workshop, novice project tutors will conduct tutoring sessions developed by trainees for additional science investigations with small groups of students who have previously completed the 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 over the course of the first three quarters of year 1.

Development Schedule: We estimate that the end-to-end process of developing tutorial dialogs and acceptable tutorial dialog systems (which will be refined further) for a complete FOSS science module will span approximately 16 weeks. During the first 8 weeks, the project staff develops and refines a set of effective dialogs. During the next 4 weeks, a system designer develops and tests the corresponding spoken dialog systems. The dialogs are then tested in Wizard of Oz sessions for an additional 4 weeks in the lab and refined for classroom use. By having two teams develop tutorial dialogs for two different FOSS modules simultaneously, it is possible to develop four dialogs for four FOSS science modules in one school year. By adding a third team of four human tutors in the 3rd and 4th quarters of year 1, with system development and testing proceeding throughout the summer, all six modules can be developed for testing in classrooms by the first or second quarter of year 2. Even if the development cycle is delayed by one or even two quarters, it will be possible to deploy and test completed spoken dialogs systems and test all treatment conditions for all modules during the second half of year 2 by simultaneous testing in classrooms—three modules in the 3rd quarter, and three modules in the 4th quarter.

Continued improvement and refinement of human and virtual tutoring dialogs will continue in year 3 (the assessment phase) based on transcriptions and analyses of tutorial dialogs recorded during small-group treatments. These dialogs will be analyzed and evaluated for their effectiveness in assessing and facilitating comprehension of science concepts incorporated into the science investigations. These dialogs will be evaluated in terms of both the quality of students' responses, and in terms of their effectiveness in identifying and addressing problems encountered by specific students based on formative assessment measures. Data from dialogs with poor outcomes will be used to improve those components of the system that caused problems, and data from dialogs with good outcomes will be used to retrain the acoustic and language models of the speech recognizer.

2.5 Assessing Feasibility and Potential of the Treatments

Overview of Assessment Plan

For ease of exposition, we assume that each participating classroom has 28 students who will participate in the research. In each participating school, students in two classrooms are about to enter the same eight-week science module—a curriculum with well-defined learning goals tied to state and district standards. For example, fourth graders in these two classrooms are about to learn about Magnetism and Electricity, with investigations entitled *The Force*, *Making Connections*, *Advanced Connections*, *Current Attractions* and *Click It*. Within each classroom, students are randomly assigned to one of three conditions: seven to Human QtA tutoring outside of class, seven to Virtual QtA tutoring outside of class with the computer system, and 14 to one of our in-class group conditions (see Table 1 and the figure in Appendix A for a graphical representation of the sampling design).

The classrooms are randomly assigned to two groups. In Group A, the 14 students assigned to the in-class condition receive standard classroom instruction (no-treatment control), while in-class students in Group B receive classroom instruction that incorporates QtA dialogs. If all four teachers assigned to Group B classrooms are willing to be trained in QtA, they will conduct this instruction; otherwise our trained QtA tutors will conduct the instruction for the experimental modules. If teachers are to be trained, they will participate in the same three-day workshop

sequence as participants in Year 1, which will again be conducted by Dr. McKeown. Of course, this example holds for third and fifth graders also.

Table 1. Sampling design

Classroom	Control: A—Normal instruction by teacher	Treatment 1: B—Classroom instruction by CSLR QtA project tutors	Treatment 2: C—Small groups with CSLR QtA project tutors	Treatment 3: D – Small groups with QtA virtual tutor
1	14		7	7
2	14		7	7
3	14		7	7
4	14		7	7
5		14	7	7
6		14	7	7
7		14	7	7
8		14	7	7
Total (N = 224)	56	56	56	56

Bulleted Description of Research Design

Third, fourth and fifth grade students in each of eight classrooms will be randomly assigned to one of four treatment conditions:

- In the *Classroom Control* condition, regular classroom teachers will present the material to groups of 12 or more students following their standard inquiry format.
- In the *QtA Classroom* condition, teachers or tutors trained in the Question the Author technique will present the material using the QtA techniques to groups equal in size to the control condition.
- In the *QtA Tutor* condition, tutors trained in the QtA technique will provide tutoring to students in groups of three or four students.
- In the *Virtual Tutor* condition, our automated tutor, developed to simulate a human QtA tutor, will present the material to students in groups of three or four students.

We assume a classroom size of approximately 28 students and will randomly assign students to groups as follows:

- 7 students receive the QtA Human Tutor condition,
- 7 students receive the Virtual Tutor condition,
- for a randomly selected half of the classrooms, the remaining group of 14 students receive the QtA Classroom Group condition, and the 14 students from the other half of the classrooms receive the business-as-usual (control) condition.

Materials and Measures

Students in all conditions will receive similar classroom learning experiences in the FOSS science modules in classrooms during science investigations that include embedded assessments and classroom discussions during which students explain their results and conclusions. We will observe classroom instruction to identify similarities and differences in classroom experiences in the different conditions. Students who remain in the classroom (Classroom Control condition) will read books that are included with the FOSS modules as supplementary learning materials while students in the other conditions will engage in QtA dialogs linked to the science investigations they are conducting.

Pre-tests, post tests and delayed post-testing will include content to be taught in the modules. These benchmark tests have been developed by Mark Wilson at UC Berkeley under a grant to the FOSS project team. In addition, the psychometric group at UC Berkeley has also developed embedded assessment measures for each investigation (i.e. each two week unit). The embedded assessments have between 8 and 12 items. They are reliable with alphas in the range of .80 to .89. Reliabilities for the pre and post are all in the mid-90's. The validity of the measures has been built up over time through a process of empirical investigation. The psychometric work draws heavily on the NRC report's "assessment triangle" which draws on a cognitive model of student learning in the domain, a model for the selection of (items) observations linked to the cognitive model and an inferential model for the interpretation of responses to items. The inferential model used here is the one parameter IRT model known as the Rasch model (see Rasch, 1960), the same model used to generate the Woodcock Johnson measures. The measures were built by generating construct maps around the key concepts. These concepts were then broken down into constructs, and then items were developed around each construct. The empirical model was iterative where items were generated and tested over time to reach fit between the cognitive and interpretive models and to further insure high validity of the test instruments.

Inter-rater reliability for subjective items on the FOSS assessments will be assured through training and monitoring of inter-rater agreement. FOSS representatives estimate that raters with some background with the modules can be trained in four hours to reach an acceptable standard of reliability. Inter-rater reliability for grading range between .79 and .89.

Learning gains will be compared across the four conditions described above by testing each student immediately before and immediately after each science module, and again two weeks or one month later. For the summative measures we will employ the hierarchical linear model to investigate the resulting individual growth curves and ascertain treatment effects.

Comparisons

Our design and measures ensure that our Control and QtA classroom conditions have enough students that individualized instruction is unlikely, while providing us the power of a within classroom design. The motivation for these conditions is to perform the following comparisons:

1. Demonstrate the magnitude of learning gains affected by classroom QtA techniques compared to standard classroom instruction (QtA Group vs. Control).
2. Demonstrate the magnitude of learning gains achieved by one-on-one tutoring vs. group administered instruction using the QtA techniques (QtA Tutor vs. QtA Group).

3. Assess how close we can come to achieving human tutoring gains by using a virtual tutor (Virtual Tutor vs. QtA Tutor, QtA Group and Control).
4. Provide insight into what aspects of the Virtual Tutor account for the difference in performance between it and human tutoring and between it and group instruction (Virtual Tutor vs. QtA Tutor, QtA Group and Control).

Hypotheses

We have three working hypotheses:

1. Based on prior research, we expect learning gains to be greater in the QtA Group condition than in the Classroom Control condition. QtA was developed for classroom use, and research has already demonstrated learning gains using the method.
2. We hypothesize that students in small groups interacting with a human tutor will have the largest learning gains. These students will receive the most effective individualized instructions by trained tutors with clearly defined learning objectives, domain knowledge and effective learning techniques. Students in these groups will produce significantly greater learning gains than student who receive QtA dialogs in classrooms. The rationale for this prediction is that students in classrooms do not receive as much individualized instruction. The teacher may call on two students to answer a question, and the students may provide good answers, but other students who were not called on may not have been able to answer the question well, and have difficulty following the ongoing dialog.
3. We predict that students who interact with the virtual tutor in small groups will have learning gains midway between students who engage in dialogs in classrooms and students who engage in dialogs in small groups with a human tutor. We do not expect the virtual tutor to have the sensitivity and flexibility of an expert human tutor, but predict that receiving individualized attention and benefits of peer discussions in small groups with the virtual tutor will produce learning gains greater than students who engage in dialogs in classrooms.

These hypotheses are well motivated, given prior research demonstrating learning gains using QtA in classrooms, and large effects, in excess of two standard deviations, following one-on-one tutoring with experts.

Assessing Outcomes

The development of the interventions proposed will be studied with both formative and summative evaluation. Qualitative and quantitative experimental studies will be conducted by teams of independent researchers.

Formative evaluation in years 1 and 2: ATLAS (Alliance for Technology, Learning and Society) evaluators will work with the CSLR development team and independently to ensure that the design of the software is usable and compatible with the curricular and practical needs of teachers (Weston, 2004). Specifically, formative evaluation of the content and usability of the application will be evaluated by a group of primary school teachers recruited from the school districts involved in this project. The teachers will participate in participatory design sessions with CSLR where they will test early versions of the software and comment on its design. Feedback will be gathered with surveys and interviews designed and field tested for this purpose.

Student feedback about tutors will be gathered through observation and think aloud sessions as interface design nears completion.

Summative evaluation: The summative evaluation in years 3 and 4 examines the effect of treatment on learner's development by treatment (and control) for those students selected for the project.

Student Characteristics: A maximum of six hundred eighty two students will participate in the study each year during years three and four. Each student will receive instruction on two FOSS modules (covering an 8 week period) and remain in the same experimental condition across modules. These students will be drawn from third, fourth and fifth grades in Boulder Valley schools that scored in the lowest one third on Colorado CSAP science tests administered to students in fifth grade. At least four different schools will participate each year; ideally, for logistical reasons, a participating school will have two third, two fourth and two fifth grade classrooms, enabling assessment to be conducted for paired classrooms at each grade level in each school. The schools participating in the study have students from lower SES backgrounds and a higher proportion of Hispanic students and English language learners than higher performing schools. Table A1 in Appendix A entitled School Demographics provides demographic data of schools that are likely to participate in the study.

Assignment and Power: The students will be randomly assigned within eight classrooms to one of four conditions during years 3 and 4 for each eight week FOSS module. Within each individual classroom, students are randomly assigned to one of three conditions: half of the students (14) remain in the classroom and receive classroom instruction and half (14) receive one-on-one tutoring using *Question the Author* dialogs. The 14 students who leave the classroom to receive individualized instruction involving QtA dialogs are randomly assigned to two conditions: half (7) receive tutoring with a human tutor, and half receive tutoring with the computer intervention. Finally, the two control conditions are alternated. In one group of classrooms, the 14 students who remain in the classroom receive standard classroom instruction (classroom control), while students in the other group receive classroom instruction that incorporates QtA dialogs. 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. Table 1 and Figure A1 in Appendix A presents the sampling design.

Power analyses use conservative estimates of effects and are based upon a one-factor Analysis of Variance design allowing for effect sizes to vary between conditions. Students experiencing human and computerized tutors are not expected to differ significantly in outcomes between these two conditions. A moderate effect size difference of .25 of a deviation unit or more is expected between the human and computer tutors (on one hand) and the control groups. Given the expectations, the power for a four group comparison with eight classrooms of 28 students each ($n = 224$) provides more than 90% power at the alpha .05 level, even if student attrition reaches 20%. Power is approximately 70% for the alpha .01 level.

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. The

BVSD Assessment and Planning group (see support letter) will provide CSAP scores (scalar values) for students in our study for CSAP items corresponding to the science concepts in the FOSS science modules.

Research setting, confounds, treatment fidelity

The two group conditions include classroom QtA instruction and a standard classroom control, each with approximately 14 students. The classroom control condition has teachers either instructing students in a manner of their own choosing or using FOSS books that present science stories related to the science investigations. Classroom control teachers will be shown the multimedia presentations that students in the three QtA dialog conditions may receive, and will have the option of presenting these materials to students. In the classroom QtA dialog condition, students will discuss science concepts in QtA dialogs managed by either trained regular classroom teachers or by our project tutors. Classroom QtA dialogs may include multimedia presentations to illustrate concepts that students are finding difficult. Presentation of materials remains constant between groups 2-4 as does time-on-task; the independent variable in the study is individual tutoring either by a computer or a human, and the comparison with the no-treatment control classes.

Because of the short duration of the treatment and the fact that treatment groups are separated physically from each other, risks of treatment contamination between students is minimal. To avoid large variations in tutoring implementation, project tutors, and presenters in the classroom control and QtA conditions will be trained to provide services in as standard a manner as possible. Researchers will observe a sample of tutors with students to learn if tutoring is carried out in an anticipated and uniform manner. Teachers in the control condition will also be interviewed and observed (with the researcher taking field notes), and instruction will be described to learn the nature of the control condition 1 and of possible variation in instructional technique between teachers. All qualitative data will be analyzed with a domain analysis.

Analysis Procedures

Scores from pre and post measures and follow-up measures will be used as dependent variables for analysis. We will use statistical techniques that leverage our capacity to randomly assign. 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 (ANOVA). 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. Variability between human tutors across different students and between classroom teachers will also be incorporated in factorial designs.

Other variables of interest incorporated into models may include person characteristics such as gender, ethnicity and language status, and differences between schools. All statistical reporting will include effect sizes, and where appropriate clustered effect size analyses.

2.6 Future Work: Scaling up Tutorial Dialogs

A more scalable (and high risk) approach to dialog design would replace the domain-specific representation being proposed with statistically-based *domain-independent shallow semantic representations*. Work towards this objective, funded under the ARDA Aquaint program, is underway at CSLR that applies computational semantic approaches to open domain question answering applications. In the past several years we have successfully applied machine learning techniques (Support Vector Machines) to the task of domain-independent shallow semantic parsing. We have used a number of domain-independent semantic role labeling schemes, including Thematic Roles, such as *Agent*, *Patient*, *Cause*, etc. In future work, proposed to the NSF, we will conduct experiments to develop more general representations and techniques as the basis for conversational interaction and answer assessment. These new technologies allow us to produce domain-independent semantic representations for comparison to the domain-dependent representations in the dialogs we will design in the proposed work. Success in this future work will allow us to build dialogs on new topics without having to write grammars specific to the domain.

3. Personnel

The partnership assembled for this project brings together cognitive psychologists, educational researchers and computer scientists with unique and complementary expertise and skills who share a passionate commitment to improving learning and student achievement in the United States by researching and developing effective and scalable scientifically-based programs. The collaborators at the Center for Spoken Language Research, the Lawrence Hall of Science, the Learning Research and Development Center and the Boulder Valley School district have each developed, implemented and assessed effective instructional programs in schools. Information about each collaborator is provided below, with additional information about their role in the project provided in the Budget Justification.

Principal Investigator, Dr. Ron Cole, is the Director of the Center for Spoken Language Research (CSLR) and a Research Professor at the University of Colorado-Boulder. He has studied speech recognition by human and machine for the past thirty years; has published over 150 articles in scientific journals and published conference proceedings. In 1990, Ron founded the Center for Spoken Language Understanding (CSLU) at the Oregon Graduate Institute. In 1998, Ron founded CSLR at the University of Colorado, Boulder. Ron is co-director of the Colorado Literacy Tutor project, a multi-agency and multi-institutional project that aims to develop virtual tutors that teach children to read and learn from text. He currently manages the Foundations to Literacy program, and four small projects funded by the NIH and NIDRR that use virtual speech therapists to improve communication skills of individuals with Parkinson Disease and aphasia. Ron will supervise and lead all project activities and serve as the principal liaison between all project personnel.

Co-PI, Dr. Barbara Wise, is a Research Associate in the Center for Spoken Language Research at the University of Colorado, as well as president of Remedies for Reading Disabilities, Inc. She has studied individual differences in reading disabilities as well as responses to intervention using “talking computers” for the last 18 years. She also has 30 years clinical experience teaching students with reading disabilities. She has written and teaches a program for teachers, Linguistic Remedies, in providing scientifically-grounded language-based interventions that take

children “beyond competence,” to application, automaticity, and transfer of newly improved skills into engaged and independent reading and writing. She has authored or co-authored at least 40 articles and chapters in scientific journals and books, and has more than 70 national and international presentations on early reading, reading disabilities, and intervention. She is Principal Investigator of a project from IES, to develop ICARE: Independent Comprehensive Adaptive Reading Evaluation with a talking and listening computer system.

Investigator, Dr. Sarel van Vuuren, is a Research Associate at the Center for Spoken Language Research at the University of Colorado, and the Head of Technical Development of the *Foundations to Literacy* learning program. A former Fulbright scholar, he has published over 30 articles in scientific journals and conference proceedings in the area of human language technologies. Over the past ten years he has led and participated in the development of several learning, assistive and informative systems, among others, the *FtL* program, an automatic speech recognition toolkit, a speaker verification system, and a virtual speech therapist for individuals with Parkinson Disease. His research areas include intelligent learning systems and agents, cognitive and machine learning, data mining and software architectures.

Co-PI, Dr. Wayne Ward is a Research Professor in the Center for Spoken Language Research at the University of Colorado, Boulder. Dr. Ward has conducted research in the area of Spoken Dialog Systems and Information Extraction by machine since 1986. He spent 12 years as a faculty member in the Computer Science department at Carnegie Mellon University before coming to the CSLR. Dr. Ward has been project leader for two different DARPA Human Language Technology program sponsored projects to build large interactive Spoken Dialog Systems. Both of these systems were top performers in yearly common evaluations. Dr. Ward developed and maintains the Phoenix system which is designed specifically 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.

Margaret G. McKeown (Investigator) is a Senior Scientist at the Learning Research and Development Center, University of Pittsburgh. A major focus of Dr. McKeown’s research has been the study of students’ comprehension from school texts. This work has included examination of textbook materials and their effect on young students’ comprehension; the design of revisions to texts based on cognitive theory; and the development of the instructional approach, Questioning the Author, to help students actively construct meaning from what they read. For the proposed project, Dr. McKeown will train tutors in Questioning the Author. She will also collaborate with project members to analyze human tutoring sessions and use that data to design and enhance a virtual tutor. Dr. McKeown will devote 25 % of her time to the project.

Finbarr Sloane is a professor at Arizona State University. His areas of expertise are Assessment, Mathematics Education, Multilevel Modelling, Longitudinal Data Analysis and Scaling of interventions. Prior to joining the faculty at Arizona State University he was a program director at the National Science Foundation's Division of Research, Evaluation, and Communication. There he oversaw a national effort to conduct research on the scaling of educational interventions in STEM disciplines. While at the NSF he provided institutional direction to Trends in International Mathematics and Science Study (TIMSS), the Board on International Comparative Studies in Education (BICSE).. Barry contributed to the proposal effort, and has agreed to work with Dr. Cole, Dr. Weston and other project staff to provide oversight and to participate in the assessment planning and analyses.

Larry Malone Mr. Malone is co-director of the Full Option Science System Project, Co-director of the Assessing Science Knowledge Project (ASK), and has been at the Lawrence Hall of Science, University of California at Berkeley, for 40 years working in science curriculum development and teacher preparation. He has been a curriculum developer and instructor for the OBIS (Outdoor Biology Instructional Strategies), HAP (Health Activities Project), GEMS Great Explorations in Math and Science), SAVI/SELPH (Science Activities for the Visually Impaired/Science Enrichment for Learners with Physical Handicaps), and FOSS K–6 and FOSS Middle School programs. Mr. Malone is a creative materials developer and designer of instructional activities, and serves as the lead writer on the FOSS materials.

Linda DeLucchi Ms. De Lucchi is co-director of the Full Optional Science System Project (FOSS K–8) and the Assessing Science Knowledge Project (ASK) at Lawrence Hall of Science, University of California at Berkeley. She has developed instructional materials in science education (FOSS), environmental education (OBIS), health education (HAP Project), and special education (SAVI/SELPH) for 32 years. In addition to curriculum development, Ms. De Lucchi has directed numerous teacher preparation projects, and has provided many tens of thousands of teacher-hours of science education inservice at the site level, district level, and national-leadership level throughout the country and abroad (Israel, Slovakia, Czech Republic, Japan, and China).

4. Facilities, Equipment and Other Resources

The primary locations involved in this project are the Center for Spoken Language Research on the Boulder Campus and the Boulder Valley School District.

Center for Spoken Language Research: The Center was established at the University of Colorado, Boulder in the Fall-1998/Spring-1999 and is part of the Institute of Cognitive Science. The Center's mission is to create systems that enable natural conversational interaction between people and machines.

The University of Colorado has provided exceptional facilities to the Center for both research and education. The University provides laboratory space, including 15 single occupancy rooms, 10 double occupancy rooms, an equipment room, a room for demonstrations and meetings, and one classroom for computer laboratory classes. In addition, the Center faculty hold appointments in academic departments (Computer Science, Psychology, Linguistics, Electrical Engineering), and therefore have access to the research and educational facilities of these departments, as well as the campus computing network, libraries, and all other available campus resources. The Center currently has 6 faculty, 9 staff, and four full-time graduate students.

The Laboratory supports both Unix and Windows operating systems, and provide powerful desktop computers for individual researchers (faculty, staff and students), and powerful data and computer servers for developing evaluation systems and for collecting data.

BVSD: The Boulder Valley School District is a 27,000 student school district with more than 20 elementary schools. Selected classrooms in these elementary schools will be used as facilities to conduct the proposed feasibility study. The technology team at CSLR and the Instructional Technology department at BVSD have worked closely together in previous research endeavors to ensure access to and connectivity between BVSD and CSLR. All classrooms in BVSD have high speed internet connectivity.

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. L, McKeown, M. G, and Kucan, L. (2002). *Bringing Words to Life: Robust Vocabulary Instruction*. New York: Guilford Press.
- 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.
- 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., Van Vuuren, S., Pellom, B., Hacıoglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W., & Yan, J. (2003). Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human-Computer Interaction, *Proceedings of the IEEE: Special Issue on Human-Computer Multimodal Interface*, 91 (9), pp. 1391-1405, Sept., 2003.

- Cole, R., Wise, B., Van Vuuren., S. (2006). How Marni teaches children to read. *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.
- Flying-Volando (2006). Available online:
<http://cslr.colorado.edu/beginweb/volando/volando.html>
- FTL Surveys (2006). Teacher and Student Histograms available online:
http://cslr.colorado.edu/beginweb/documents/ftl_student_teacher_surveys.html
- Graesser, A.C. and Person, N.K. (1994) Question Asking During Tutoring. *American Educational Research Journal*, 31 (1), 104-137.
- 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.
- Hagen, A., Pellom, B., Cole, R. (2003) "Children's Speech Recognition with Application to Interactive Books and Tutors", in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, St. Thomas, USA, Dec., 2003*.
- Hagen, A., Pellom, B., van Vuuren, S., Cole, R. (2004) "Advances in Children's Speech Recognition within an Interactive Literacy Tutor", in *HLT NAACL 2004, Boston, May, 2004*.
- 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.

- Kintsch, W. (1988) "The role of knowledge in discourse comprehension: A construction-integration model", in *Psychological Review*, 95, 163-182.
- Kintsch, W. (1998) "Comprehension: A paradigm for cognition", Cambridge, England. Cambridge University Press.
- Kintsch, E. (2005). Comprehension theory as a guide for the design of thoughtful questions. *Topics in Language Disorders*, 25 (1), 51-64.
- Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B., & Boghal, R. (1997). The persona effect: Affective impact of animated pedagogical agents. In *Proceeding sof CHI 97 human factors in computer systems*, pp. 359-366. New York: Association for Computing Machinery.
- 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.
- Meichenbaum, D. and Biemiller, A. (1998) *Nurturing independent learners: Helping students take charge of their learning*. Cambridge, MA: Brookline.
- 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.
- Rose, C. P., Moore, J. D., VanLehn, K., Allbritton, D. (2001). A Comparative Evaluation of Socratic versus Didactic Tutoring, Proceedings of the Cognitive Sciences Society (poster).
- 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.
- Snow, C. (2002). Reading for Understanding: Toward A R & D Program in Reading Comprehension, Rand Education.
- 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.

- Sweet, A. P. and Snow, C. E. (Eds.). (2003) Rethinking reading comprehension. New York: Guilford Press.
- Taylor, T., Black A., and Caley, R. (1998). "The Architecture of the Festival Speech Synthesis System", in *Proc. Third ESCA Workshop in Speech Synthesis*, Blue Mountains, Australia, pp. 147-151. Available: <http://www.cstr.ed.ac.uk/pubs>.
- Topping, K., & Whitley, M. (1990) "Participant evaluation of parent-tutored and peer-tutored projects in reading", in *Educational Research*, 32(1), 14-32.
- VanLehn, K. & Graesser, A. C. (2002). 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. (2002). In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Intelligent Tutoring Systems 2002* (pp. 367-376). Berlin, Germany: Springer.
- 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

APPENDIX A

Figure A1. Research design.

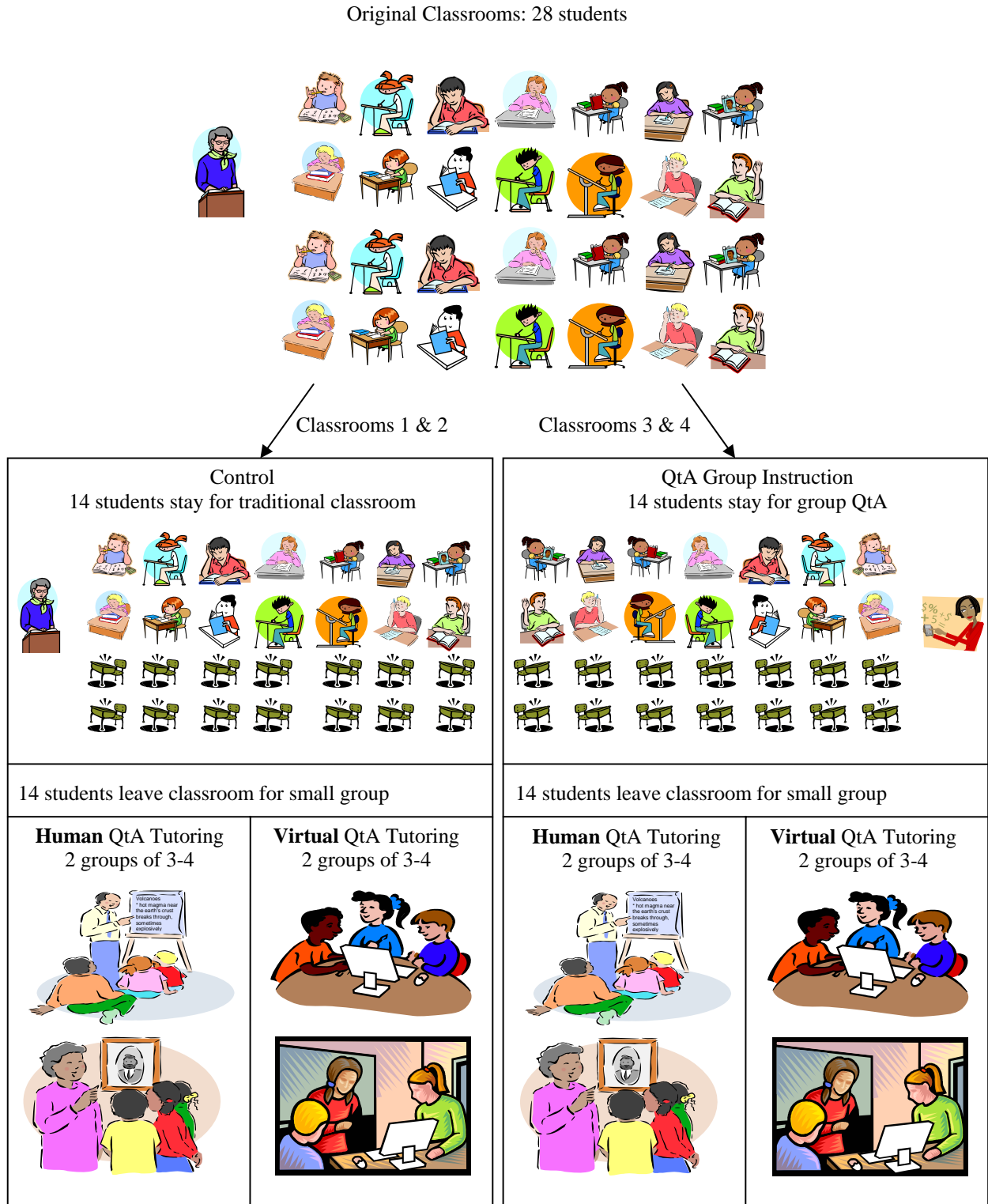


Table A1. School demographics (as percent (%) per category)

Category	School									
	Birch	Columbine	Creekside	Emerald	Lafayette	Pioneer	Ryan	Sanchez	University Hill	Whittier
Female	50	48	45	51	48	49	52	55	53	55
Male	50	52	55	49	52	51	48	45	47	45
African-American	2	1	4	1	3	1	3	3	1	5
American Indian	1	1	3	0	1	1	2	0	1	0
Asian	3	2	12	10	5	1	7	3	2	10
Caucasian	83	14	52	50	71	37	71	30	27	68
Hispanic	11	82	30	39	21	59	17	64	70	17
Free Lunch	13	1	38	43	25	40	21	65	56	33
Reduced Lunch	3	6	7	9	4	8	6	8	7	3
English Language Learners	1	69	24	24	11	38	5	25	48	15
Special Education	9	11	19	16	18	6	14	15	8	11

APPENDIX B

FOSS tables for each Science Module

Table B1. Measurement (3rd grade)

Colorado Science Standard 1: Scientific Investigations

Investigation	Main Concepts (from FOSS Module Summary)
“The First Straw”	<p>A measurement standard is a unit agreed upon and used by a large number of people.</p> <p>A standard unit of measure is necessary for consistency in communication between people.</p> <p>A meter (m) is the basic unit of linear measurement in the metric system.</p> <p>A centimeter (cm) is 1/100 of a meter.</p> <p>A kilometer (km) is 1000 meters.</p> <p>An estimate is an approximate value.</p>
“Weight Watching”	<p>A measurement standard is a unit agreed upon and used by a large number of people.</p> <p>The gram (g) is the standard unit of measure used to determine mass in the metric system. (1 g equals 1 cc or 1ml of water at 4°C.)</p> <p>Mass is how much of something there is.</p> <p>A kilogram is 1000 g, or the same as the mass of 1 liter of water.</p>
“Take Me To Your Liter”	<p>A measurement standard is a unit agreed upon and used by a large number of people.</p> <p>Volume is the three-dimensional space occupied by something.</p> <p>Capacity is the volume of fluid a container can hold when full.</p> <p>The liter (L) is the standard for measuring fluid volume in the metric system.</p> <p>One liter is equal to 1000 milliliters (ml).</p>
“The Third Degree”	<p>A measurement standard is a unit agreed upon and used by a large number of people.</p> <p>Degrees Celsius (°C) is the unit scientists generally use to measure temperature.</p> <p>The Celsius temperature scale is based on the freezing (0°C) and boiling (100°C) points of water.</p> <p>A change in temperature is a result of addition or subtraction of heat.</p>

The Measurement unit immerses students in the metric system of measurement and attempts to build students’ ability to “think metrically”. Students are also expected to develop a solid conceptual understanding of mass and volume.

Table B2. Physics of Sound (3rd grade)

Colorado Science Standard 1: Scientific Investigations

Colorado Science Standard 2: Physical Science

Investigation	Main Concepts (from FOSS Module Summary)
“Dropping In”	Sounds have identifiable properties. Objects can be identified by the sound they make when dropped. The identifiable properties of sounds can convey information. Sustained sound is caused by vibrations. Sound requires a source and a receiver. The intensity of the vibration determines the volume.
“Good Vibrations”	Sound originates from vibrating sources. Pitch is how high or low a sound is. Differences in pitch are caused by differences in the rate at which objects vibrate. Pitch can be changed by changing the length or tension of the object vibrating at the sound source.
“How Sound Travels”	Sound vibrations need a medium to travel. Sound travels through solids, such as wood. Sound travels through liquids, such as water. Sound travels through gases, such as air. Sound energy can be directed with reflective tubes and megaphones. Our outer ears are designed to gather sound energy.
“Sound Challenges”	Several variables affect pitch, including size (length) and tension of the vibrating object at the sound source. Sound can be directed through air, water, or solids to the sound receivers. The medium that sound passes through affects its volume and the distance over which it can be heard.

This unit is challenging for third graders. For many students, this is their first exposure to the concept that sound is caused by vibrations. In addition, students are introduced to the idea that sound requires a medium through to travel.

Table B3. Water (4th grade)

Colorado Science Standard 1: Scientific Investigations

Colorado Science Standard 4: Earth Science

Investigation	Main Concepts (from FOSS Module Summary)
“Water Observations”	Water has several observable properties, including transparency, shapelessness, and movement or flow. Water beads up on some materials and is absorbed by other materials. Surface tension is the skinlike surface of water that pulls it together into the smallest possible volume. Water flows downhill.
“Hot Water, Cold Water”	Water contracts when heat is taken away. Cold water is denser than warm water. Water is densest at 4 °C. Ice is less dense than liquid water. A solid has a definite volume and shape; a liquid has only definite volume.
“Water Vapor”	Evaporation is the process by which liquid water changes into water vapor, a gas. Temperature affects the rate of evaporation. The surface area of a liquid affects the rate of evaporation. Condensation occurs when water vapor touches a cool surface and changes into liquid. Evaporation and condensation contribute to the movement of water through the water cycle.
“Waterworks”	Water flows more easily through some earth materials than through others. Flowing water can be used to do work. Water contains different materials that affect its quality. Evaporation can be used to detect materials dissolved in water.

This unit introduces the properties of water, the water cycle, and the concepts of phase change and density. These concepts are foundational to an understanding of the role water plays in the natural world, from living things to basic chemistry, to weather. Students may benefit especially from coaching in how to verbalize the relationship between density and the phenomenon of sinking and floating.

Table B4. Magnetism and Electricity (4th grade)

Colorado Science Standard 1: Scientific Investigations

Colorado Science Standard 2: Physical Science

Investigation	Main Concepts (from FOSS Module Summary)
“The Force”	Magnets stick to metal objects made of iron. Magnetic interactions are caused by the magnetic force. Magnets display forces of attraction and repulsion that decrease with distance. Magnetism can be induced in a piece of steel that is close to or touching a magnet.
“Making Connections”	Electricity flows through pathways called circuits. A switch is a device used to open and close circuits. An open circuit is an incomplete electric pathway; a closed circuit is a complete pathway. Materials that allow electricity to flow are conductors; those that do not are insulators.
“Advanced Connections”	A circuit with only one pathway for current flow is a series circuit. (Components “share” the electric energy.) A circuit with two or more pathways for current flow is a parallel circuit. (Components each have a direct pathway to the energy source.)
“Current Attractions”	A core of iron or steel becomes an electromagnet when electricity flows through a coil of insulated wire surrounding it. There are a number of ways to change the strength of an electromagnet, including changing the number of winds of wire around the core.
“Click It”	An electromagnet placed in a complete circuit can be used to make a telegraph. A switch can serve as a key in a telegraph system. A code is a symbolic system used for communication. Technology is the application of science.

Even adults struggle to put into words the relationships between magnetism and electricity. Because this unit deals with unseen forces, multiple representations of those forces may help them develop a more solid understanding.

Table B5. Variables (5th grade)

Colorado Science Standard 1: Scientific Investigations

Colorado Science Standard 2: Physical Science

Investigation	Main Concepts (from FOSS Module Summary)
“Swingers”	A variable is anything that you can change in an experiment that might affect the outcome. In a controlled experiment only one variable is changed, and the results are compared to a standard. The length of a pendulum determines the number of swings in a unit of time.
“Lifeboats”	A variable is anything that you can change in an experiment that might affect the outcome. In a controlled experiment, only one variable is changed, and the results are compared to a standard. Capacity is the maximum volume of fluid a container can hold.
“Plane Sense”	A variable is anything that you can change in an experiment that might affect the outcome. In a controlled experiment, the experimental variable is changed incrementally to see how it affects the outcome. A system is a set of related objects that can be studied in isolation.
“Flippers”	A variable is anything that you can change in an experiment that might affect the outcome. In a controlled experiment, the experimental variable is changed incrementally to see how it affects the outcome. A system is a set of related objects that can be studied in isolation.

Variables is a challenging unit for 5th graders, but critical to their development of the abilities to do scientific inquiry. According the National Science Education Standards, even middle school students may have difficulty identifying and controlling variables in an experiment. Students may also benefit from coaching in how to verbalize relationships between variables.

Table B6. Environments (5th grade)

Colorado Science Standard 1: Scientific Investigations

Colorado Science Standard 3: Life Science

Investigation	Main Concepts (from FOSS Module Summary)
“Terrestrial Environments”	Everything that surrounds an organism makes up the organism’s environment. An environmental factor is one part of an environment. It can be living or nonliving. Terrestrial environments include both living and nonliving factors. A relationship exists between environmental factors and how well organisms grow.
“Bugs and Beetles”	Designing an investigation involves controlling the variables so that the effect of one factor can be observed. Each organism has a set of preferred environmental conditions.
“Water Tolerance”	Organisms have ranges of tolerance for environmental factors. Organisms have specific requirements for successful growth, development, and reproduction. Optimum conditions are those most favorable to an organism’s survival, growth and reproduction.
“Brine Shrimp Hatching”	Environment is everything that surrounds and influences an organism. An environmental factor is one part of an environment. It can be living or nonliving. Optimum is the condition most favorable to growth, development and reproduction of an organism. Range of tolerance is the conditions in which an organism can survive. Range lies between the high and low extremes of tolerance for an environmental factor.
“Salt of the Earth”	Environment is everything that surrounds and influences an organism. An environmental factor is one part of an environment. It can be living or nonliving. Optimum is the condition or degree of an environmental factor that is most favorable to growth, development and reproduction of an organism. Organisms have ranges of tolerance for environmental factors.

This unit introduces concepts related to populations of organisms and the unique environments in which they live. Some students may have difficulty thinking about plants or animals in terms of populations rather than as individuals. Likewise, students may be unaccustomed to thinking about environments as having characteristics that affect how well they are suited to providing habitat for different organisms. An especially difficult concept in this unit is introduced in “Aquatic Environments”. In this activity students are expected to build an understanding that respiration by animals causes water to become more acidic due to the accumulation of CO₂.

Table B7. Full option science system.

(From page 58 of the Embedded Assessment Portfolio, Ask Project, FOSS).

 ASK—MAGNETISM AND ELECTRICITY ASSESSMENT			
	EMBEDDED		BENCHMARK
	Assessment format	Focus Constructs	Survey
INVESTIGATION 1			
Part 1 Investigating Magnets...	Teacher Observation <i>Magnet Interactions</i> , No. 2	ME1-ir ME1-ar	
Part 2 Investigating More...	More Mag Observations, No. 3	ME1-sm ME1-tm	Inquiry Scenario 1
Part 3 Breaking the Force	Teacher Observation <i>The Force—Graph</i> , No. 6	ME1-sm IN1-vp IN2-cg IN3-pd	
Part 4 Detecting the Force...	Teacher Observation	ME1-sm ME1-ir	Investigation 1 I-Check
INVESTIGATION 2			
Part 1 Lighting a Bulb	Teacher Observation	ME2-pc ME2-fc ME2-ef	
Part 2 Making a Motor Run	Teacher Observation <i>Response Sheet—Inv 2</i> , No. 10	ME2-pc IN2-cg	
Part 3 Finding Conductors...	<i>Conductors and Insulators</i> , No. 11	ME2-pc IN2-cg	
Part 4 Investigating...Circuits	Teacher Observation	ME2-pc	Investigation 2 I-Check
INVESTIGATION 3			
Part 1 Building Series Circuits	Teacher Observation	ME2-sp	
Part 2 Building Parallel Circuits	Teacher Observation	ME2-sp IN2-cg	Inquiry Scenario 2
Part 3 Solving the...Problem	<i>Recommend... to the Board</i> , No. 15	ME2-sp ME2-pc ME2-ef ME2-fc	Investigation 3 I-Check
INVESTIGATION 4			
Part 1 Building an Electromagnet	Teacher Observation	ME3-wr ME3-ce ME3-em	
Part 2 Changing No. of Winds	Teacher Observation <i>Changing...Winds Graph</i> , No. 17 <i>Response Sheet—Inv 4</i> , No.18	ME3-st IN2-cg IN3-pd	
Part 3 Investigating More...	<i>Electromagnet Investigations</i> , No. 19	ME3-st ME3-ce ME3-em	Investigation 4 I-Check
INVESTIGATION 5			
Part1 Reinventing the Telegraph	Teacher Observation	ME2, ME3	
Part 2 Sending Messages...	Teacher Observation	ME2, ME3	
Part 3 Choosing Your Own...	Teacher Observation <i>Choosing Your Own Inv...</i> , No. 22	ME1, ME2, ME3 IN1, IN2, IN3	
			Posttest
			Calibration Study (one week later)