

COLLABORATIVE RESEARCH:
IMPROVING SCIENCE LEARNING IN INQUIRY-
BASED PROGRAMS

Wayne Ward & Ron Cole
Boulder Language Technologies

Contents

PROJECT SUMMARY	3
Five years of MyST Research and Development	5
Organization of the Report	7
1. My Science Tutor—Spoken Dialog System (MyST- SDS)	8
The Intervention – MyST-SDS.....	11
Types and Uses of Media in MyST	13
Student Interface	15
System Operation.....	15
Stages of Tutorial Development in MyST	17
MyST Development Sequence	17
Human Tutoring.....	17
Wizard of Oz (WOZ) system and data collection.....	18
System Development	20
Data Collection and Corpus Development	20
Speech Files	21
Log files	21
Concept Annotation	21
MyST System Component Evaluations.....	22
Automatic Speech Recognition Performance	22
Concept Accuracy.....	23
Summative Evaluation of MyST-SLS	23
Is it Feasible to Integrate MyST-SLS into Elementary Science Curricula?	25
2. MyST-MP&D	28
Quantitative Results	32
MyST-MP&D Summative Evaluation.....	32
Comparisons with treatment groups from 2010-2011	34
APPENDIX A.....	49
MyST’s Theoretical and Empirical Foundations.....	49
1. Sociocultural Perspectives on Learning.....	49
2. Empirical Foundations of MyST Dialogs.....	50

PROJECT SUMMARY

My Science Tutor (MyST) is an intelligent tutoring system designed to improve children's excitement about and motivation to learn science, their ability to reason and talk about science, and their science achievement. MyST features conversational interaction with a lifelike computer character, the virtual tutor Marni, in rich multimedia environments. Conversations with Marni are designed to scaffold learning so that children can explain the science presented in illustrations, animations or interactive simulations.

The specific objectives of the proposed MyST project were:

1. Develop, through iterative design-test-and refine cycles, a set of tutorial dialogs in which children converse with a virtual science in 4 different areas of elementary school science.
2. Create a corpus to support training and evaluation of the MyST system components—speech recognition, natural language understanding, dialog modeling, speech and language generation by the virtual tutor, and presentation of media within dialogs.
3. Conduct a summative evaluation of MyST to assess the feasibility of integrating the program into classroom science instruction, and its ability to improve students' science understanding.

All of these project objectives were accomplished. Summative evaluations of two versions of the MyST system produced positive user experiences and significant learning outcomes, equivalent to human tutoring, and indicated the feasibility of integrating MyST into real world educational environments. Students were fully engaged in tutorial dialogs, and reported that they were more motivated to learn science after working with Marni. Teachers reported that they believed their students benefitted from MyST and that the tutorial dialogs aligned well with their learning goals.

Two MyST Systems: MyST-SDS and MyST-MP&D

Development of MyST was supported by two research grants awarded to BLT in 2007; the NSF DRK-12 grant, which spanned six years, and included collaboration with researchers at University of Colorado, and a four year grant from the Institute for Education Sciences' Cognition and Student Learning Program (IES-CASL). During the first three years of the project, the focus of both projects was the development of a set of sixteen tutorial dialogs for four areas of science. These 15 to 20 minute dialogs were designed to enable 3rd, 4th and 5th graders to learn to construct science explanations indicating a deep understanding of science concepts. During the fourth year of the project, we conducted the first summative evaluation of MyST. We refer to this initial version of MyST as **MyST-SDS** (Spoken Dialog System), since students spent nearly the entire tutoring session conversing with Marni. Transcriptions of MyST dialogs indicated that students and Marni were talking about 70% of the time during tutorial dialogs, and that students and Marni spoke about the same amount of time (~6 minutes each) during an average dialog of 15 to 20 minutes.

In the fourth and fifth year of the project, we developed and evaluated **MyST-MP&D** (Multimedia Presentations & Dialogs). As the name implies, MyST-MP&D combined narrated multimedia science explanations with tutorial dialogs that assessed students' understandings and interacted with them to construct accurate answers and explanations. A second major difference between the two systems is that MyST-MP&D supported both one-on-one tutoring sessions and

tutoring sessions with groups of 3 students. In small group sessions, students were encouraged to discuss answers to Marni's questions before one of the students provided an answer.

Major Outcomes of the NSF DRK-12 grant

MyST provides strong evidence for a new generation of intelligent tutoring systems. Our review of the scientific literature indicates that MyST is the first intelligent tutoring system to engage children in spoken dialogs with a virtual tutor to improve science learning. Prior to MyST, it was unknown whether human language technologies were capable of supporting spoken dialogs between children and intelligent agents. Analyses of MyST dialogs indicated that a) between 75% - 80% of the time during 15 to 20 minute dialog sessions either Marni was asking students questions or students were explaining science to Marni; b) students and Marni spent about the same amount of time talking during dialog sessions, around 6 minutes each; and c) the vast majority of students were fully engaged through each dialog session. Summative evaluation of two different version of MyST indicated that students who used MyST achieved learning gains equivalent to human tutoring, with moderate effect sizes—averaging about .5 standard deviation improvement relative to students who did not receive tutoring.

IES Replication and Efficacy Study: The successful outcomes of the MyST project resulted in the IES funding a 4-year grant to replicate and demonstrate the efficacy of MyST with a broad and diverse population of students (40 classrooms each over three years). The grant received an outstanding score by the review panel, and was one of relatively few Goal 3 grants awarded by the IES because of across-the-board government budget cuts. As of this writing (January, 2014), fourth and fifth grade students are interacting with Marni in three school districts in Colorado. By the conclusion of the project, approximately 1,200 4th and 5th grade students will have interacted with Marni for 7 to 14 hours in five areas of science. The study will produce a massive amount of speech data (one half year of continuous speech) that can be mined and analyzed to understand children's dialogs and improve the performance of the underlying speech and language technologies.

Conversations About Science Using Media (CASUM): The DRK-12 grant supported development and pilot testing of a classroom intervention in which teachers managed classroom conversations in which students learned to construct explanations of science presented in Flash animations. The CASUM intervention provided professional development to teachers who learned to a) control Flash animations (developed during the MyST-MP&D project) that presented science phenomena and systems, b) stop the presentation at strategic points, and c) ask students open-ended questions that stimulated them to share and build on each other's ideas to construct science explanations. CASUM was tested in 18 classrooms with English learners with low English language proficiency, and special needs students. Teachers' reports provided strong evidence for the feasibility of implementing CASUM dialogs in classrooms.

GROMINDS was funded by a supplement to the DRK-12 grant that supported collaboration between researchers at Boulder Language Technologies, Southern Methodist University in the U.S., and researchers at the University of Jyvaskyla in Finland, as part of an NSF Science Across Virtual Institutions (SAVI) program. The collaboration resulted in enhanced English, Spanish and Finnish versions of *MindStars Books*, are designed to help children learn science through narrated multimedia science explanations, followed by question-answer dialogs about the science, and to help them learn to read grade-level science texts accurately and fluently. The study also developed American English and American Spanish versions of *Graphogame*,

developed by our colleagues at University of Jyväskylä, a computer game that has been shown to help children acquire sound-letter correspondences and word-level automaticity, skills that are foundational to word recognition and fluent reading. MindStars Books were tested in K-2 classrooms in Colorado, and American English and Spanish versions of Graphogame were tested in second grade classrooms in Texas with both English-only speakers and English learners. The initial pilot studies indicated that feasibility and promise of the programs, and their potential for future use in classrooms for young learners worldwide. In spring of 2014, Dr. Doris Baker will incorporate MindStars Books into a Masters-level course on bilingual education for in-service teachers at SMU. SMU awarded a grant of \$10,000 to Dr. Baker to provide computers for the course. Teachers will develop their own books, integrate them into their classroom science activities, and assess students' learning using the books.

Polish Classroom Interventions Inspired by MyST & CASUM: Our team at BLT collaborated with researchers in the Center for Speech and Language Processing at the Adam Mickiewicz University (AMU) in Poznan Poland. CSLP has received two major EU grants (Prof. Katarzyna Dziubalska-Kolaczyk, PI) to develop and evaluate classroom interventions in which teachers engaged children in conversations about science shown in Flash and HTML5 animations, followed by computer-based tutoring sessions. These projects were inspired by the CASUM and MyST interventions supported by the DRK-12 grant and benefitted from active collaboration with BLT researchers and ETOS developers, and media developed at BLT. The ETOS project was conducted in Polish primary and junior high schools as an after school program, with teachers conducting CASUM dialogs which were followed by computer-based tutoring sessions. Summative evaluation revealed significant learning gains, and excellent experiences by teachers and students (<http://wa.amu.edu.pl/e-nauczyciel/> - Polish only). The Tablit project, currently being piloted in Polish kindergartens, is extremely ambitious—the project team has developed an entire inquiry-based preschool and kindergarten science curriculum composed of nine four-week science modules. Each module includes hands-on activities, CASUM conversations, computer-based tutoring, group projects, and integration of music and art into work products. If the curriculum receives positive reviews by teachers, and produces significant learning gains relative to control classrooms, schools throughout Poland will be able to choose to use the curriculum.

Five years of MyST Research and Development

The main focus of the DRK-12 grant was development and summative evaluation of two versions of My Science Tutor. Here we present a brief summary of these activities. Detailed descriptions of the systems and outcomes of the evaluation are provided below.

Year 1: School Year 2007-2008

During the first year of the project, we developed and tested 16 tutorial dialogs for each of two FOSS modules: Magnetism & Electricity (M&E) and Measurement (MMNT). All tutoring sessions involved face-to-face tutoring with a project tutor trained to conduct tutorial science dialogs using principles of “Questioning the Author” (QtA). As illustrations and animations were developed, tutors used laptops to present media during the tutorial dialogs. Individual students averaged approximately 12 tutoring sessions with human tutors. The main outcome of this phase of the research was the development of tutorial dialogs that incorporated media aligned to science concepts and learning objectives in the first two FOSS modules we developed.

The dialogs were recorded and transcribed and analyzed to improve the dialog moves and inform tutors on best practices using QtA.

We worked with *147* students in 11 different classrooms in 4 Boulder Valley School District (BVSD) elementary schools. 77 students worked with M&E (Magnetism and Electricity) and 70 students worked with MMNT (Measurement).

Altogether, 147 students were tutored during 1,764 individual sessions.

Year 2: School Year 2008-2009

During the second year of the project we continued face-to-face tutoring and also initiated “Wizard of Oz” (WoZ) sessions, in which human tutors monitored and were able to control system behaviors, unbeknownst to students. Human tutors were present during WoZ sessions in year two. They worked on a laptop at the same table as the student, but the student could not see the human tutor’s computer screen. Human tutors were able to listen to both Marni’s and the students’ speech, and could see what was displayed on the students’ screens. Tutors could hear what students were saying (via headphones) and knew what was being shown on the student’s computer screen. Tutors were presented with dialog moves and visuals that the system suggested for use, which they could approve or override.

We worked with *186* students this year, in 14 different classrooms, in 5 different BVSD elementary schools. 102 students worked with M&E, 72 students worked with MMNT, and 12 students worked with VAR (Variables).

Altogether, 186 students participated in a total of 2,232 individual sessions.

Year 3: School Year 2009-2010

Year 3: This school year focused on “Wizard of Oz” (WoZ) tutoring sessions. Children interacted with Marni in their schools while remote project tutors (at Boulder Language Technologies) viewed the students’ computer screens and listened to their dialogs with Marni. The human tutors viewed the dialog moves and media the system was about to produce, which they could approve or override.

213 Students were tutored in 18 different classrooms in 7 different BVSD elementary schools. 50 students worked with M&E, 83 students worked with MMNT, 44 students worked with VAR (Variables Module) and 36 students worked with H2O (Water). On average, individual students received about 12 tutoring sessions.

Altogether, 213 Students participated in a total of 2,508 individual tutoring sessions.

Year 4: School Year 2010-2011

During the first evaluation of MyST (MyST-SDS), 438 students were tutored: 219 students interacted with Marni independently, and 219 received tutoring with human tutors in small groups. Students used MyST in 16 different classrooms, in 7 different BVSD elementary schools. 49 students worked with M&E, 106 students worked with MMNT, 33 students worked with VAR (Variables), and 31 students worked with H2O (Water).

The 439 students participated in a total of 4672 individual tutoring sessions.

Year 5: School Year 2011-2012

The second evaluation of MyST (MyST-MP&D) included *183* students in 13 different

classrooms in 4 different BVSD elementary schools. This version of MyST used two FOSS modules, M&E and MMNT. 100 students worked with M&E and 83 students worked with MMNT. Students interacted with Marni either one-on-one or in small groups consisting of three students. Students in small groups were encouraged to discuss answers to Marni's questions before one student responded. 114 students worked in groups of 3 students and 69 students interacted with Marni one-on-one.

Altogether, 183 students participated in 1712 tutoring sessions (608 group sessions, 1104 individual sessions).

***Summary:* Over the 5 years of the MyST project, approximately 1,168 students were tutored by human tutors, during Wizard of Oz sessions, and by the virtual tutor Marni. Altogether, there were approximately 12,800 tutoring sessions.**

All tutoring sessions were recorded and transcribed. On average, children produced about 6 minutes of speech during tutoring sessions. Across all sessions, we collected over 1000 hours, or 42 full days, of children's speech. The transcribed speech data were used to train and evaluate the performance of the speech recognizer, and to evaluate the performance of the MyST system in recognizing concepts children expressed during their dialogs with Marni.

Organization of the Report

The report is organized into 4 sections.

Section 1 describes development and evaluation of the initial *MyST Spoken Dialog System*, MyST-SDS, which compared tutoring using MyST-SDS versus human tutoring.

Section 2 describes *MyST-Multimedia Presentations & Dialogs*, MyST-MP&D, and its evaluation with both individual students and small groups of students.

Section 3 describes *CASUM*, a teacher-controlled classroom intervention piloted in two successive summers in a science camp for English learners and special needs students.

Section 4 describes *GROMINDS*, an international collaboration between researchers at BLT, SMU and University of Jyväskylä in the context of an NSF SAVI project.

Appendices provide supplementary information, including an overview of theories and empirical research that informed the design of MyST, CASUM and MindStars books.

1. My Science Tutor—Spoken Dialog System (MyST- SDS)

The MyST Vision: A Virtual Tutor for Every Child

Our vision when developing MyST was to create a safe, comfortable and stimulating learning environment *in which all children could learn to engage in scientific discourse and construct explanations that demonstrate a deep understanding of science.* MyST is based on the assumption that all children can learn science, regardless of their race, ethnicity, socioeconomic status, linguistic abilities or cultural background. MyST attempts to optimize science learning by keeping children within their zone of proximal development (discussed below), where they can build on their prior knowledge and language skills to learn and master prerequisite vocabulary and concepts, and receive the scaffolding they need to construct new knowledge and communicate their understandings to a virtual tutor.

One of our greatest challenges in developing MyST was to enable children with vastly different vocabularies, discourse skills, cultural perspectives and prior experience to engage in scientific discourse. *Our approach was to recognize what children were trying to communicate using their available language skills, and build on their existing knowledge by scaffolding learning to help them understand and use scientific vocabulary to explain science.*

How was this accomplished? We analyzed children's speech collected during the MyST development process to interpret what they were trying to communicate when talking about science. These data represent the many different ways that different children talk about science in the classroom. We collected and transcribed data from over 1000 students in over 10,000 tutoring sessions during the development phase, including many English learners, during human tutoring and Wizard of Oz tutoring sessions. We were able to use the transcribed speech data to develop grammars to represent how children expressed their science understandings through their speech. These grammars were used to represent the concepts students were trying to express in the MyST spoken dialog system. During spoken dialogs with students, the virtual tutor Marni then *rephrased students' answers while modeling the correct use of vocabulary terms* that students had not yet included in their answers. Thus, Marni continually modeled the appropriate use of scientific discourse based on the ideas the student had expressed in their speech. This was followed by an open-ended question, which also modeled scientific discourse and was designed to scaffold learning and stimulate the child to construct new knowledge.

To achieve this goal, MyST continuously assessed students' science understandings by analyzing their explanations to determine which concepts they had addressed, and which concepts they had had not yet communicated, and might not know. Based on these analyses, the system selected Marni's prompts and media presentations to scaffold learning and stimulate children to build on their prior knowledge, reason about the science in the media, and construct accurate answers. We learned that these dialog moves motivated students and focused their attention as they worked with Marni to construct increasingly sophisticated and accurate explanations.

Why MyST Worked

We identified several key factors that led to successful outcomes in the MyST project. These included a) MyST dialogs' precise alignment to classroom science instruction, b. The design of the spoken dialogs, which were structured to optimize learning using established tutoring strategies inspired by sociocultural views of learning, and were modeled on the spoken behaviors of expert tutors and children during thousands of tutoring sessions, and, c) Dialogs involved

students constructing explanations about science presented in illustrations, animations and interactive simulations, which enabled children to visualize the science they were talking about, leading to rich multimodal representations and mental models of their science knowledge. We briefly discuss each of these factors.

MyST was aligned with classroom science instruction. One of the most important decisions we made, articulated in the MyST proposal to the DRK-12 program, was to align MyST dialogs to the Boulder Valley School Districts' science curriculum, which uses the Full Option Science System (FOSS). The main reason that teachers viewed MyST as an important and valuable resource that improved students' motivation and science learning was *each MyST dialog reinforced classroom science instruction*. Specifically, MyST dialogs helped individual students reason and talk about the science they encountered in the 16 classroom science investigations and associated instructional activities in each FOSS science module. These classroom activities typically included a) having students first learn to understand and use vocabulary associated with the materials and concepts encountered in investigations, b) conducting science investigations in small groups, c) writing and drawing in their science notebooks (e.g., making predictions before an investigation, summarizing their observations following it), and d) participating in teacher-led "making meaning" sessions in which the teacher led discussions to help students make sense of their experiences and observations. The MyST dialogs were designed to help students achieve a deeper understanding of the science by explaining it to Marni.

The importance of the close alignment between the MyST tutorial dialogs and classroom instruction cannot be overemphasized. The knowledge that students acquired during classroom activities, including familiarity with the science vocabulary, their hands-on experience conducting investigations, and their entries in science notebooks, provided them with substantial foundational knowledge that helped them to engage with Marni and converse with her to construct science explanations.

MyST dialogs modeled the performance of expert human tutors. Our goal in designing MyST dialogs was to have the virtual tutor Marni provide the same level of individualized and adaptive instruction as an expert human tutor. In fact, *all of Marni's tutorial dialogs sessions were modeled on dialogs between expert tutors and children*. The expert tutors received training on the learning goals and students' challenges for each science topic, and received training and feedback on their tutoring performance (from Margaret McKeown, co-developer of Questioning the Author, the dialog strategy used in MyST). Tutors thus became highly proficient at conducting tutorial dialogs in which they modeled scientific discourse, scaffolded learning through questions and presentation of media, and provided formative feedback and positive reinforcement to students contingent on the quality of their explanations.

Each MyST tutorial dialog session was organized as a sequence of mini-tutorials. The goal of Marni's dialog moves in each mini-tutorial was to have students master the vocabulary and targeted concepts learned in each one, and build on these concepts to construct a complete and accurate explanation of the science. The dialog moves were designed to help students' build on their current understandings, reason about the science, and construct explanations that communicated their new knowledge. Marni's dialog moves—strongly influenced by Vygotsky's writings and by empirical evidence on effective tutoring strategies, are described in detail in Appendix 2.

The virtual tutor Marni also played a key role in engaging and motivating students. Survey results, presented below, indicate that over 95% of students thought Marni was an engaging and effective tutor. To a large extent, Marni represented the face, voice and personality of MyST. Marni's voice was recorded by an expert tutor, who understood the purpose of each question, and the importance of formative feedback; each utterance was therefore produced with appropriate prosody. These recordings caused Marni to take on the "personality" of the human tutor.

MyST dialogs used media to help students visualize, understand, and explain science. MyST tutorial dialog sessions incorporated illustrations, animations, and interactive simulations. These enabled students to establish joint attention with the virtual tutor, visualize the science, and focus the discussion on the science presented in the media. The integration of media into MyST dialogs was based on established principles of multimedia learning, discussed below, and in Appendix 2.

MyST System Development

During the first 3 years of the project, tutorial dialogs were developed and tested for four FOSS Modules. Each module contains 4 Investigations (e.g., Magnetism, Serial Circuits, Parallel Circuits, Electromagnetism), with each Investigation divided into 4 tutorial sessions. These tutorial sessions were aligned to classroom science investigations with the kit-based FOSS science program, discussed below. A total of 64 tutorial sessions were developed and tested. During the 4th and 5th years of the project, two versions of the MyST system were developed: MyST-Spoken Dialog System (SLS) featured one-on-one spoken dialogs between 3rd, 4th and 5th with the virtual tutor Marni about science presented in media; MyST-Multimedia Presentations & Dialogs (MP&D) combined multimedia presentations of science with question-answer dialogs, and investigated both one-on-one and small group tutoring with Marni. The process of developing the two MyST systems is described in some detail below. Additional detail can be found in two journal applications (Ward et al., 2013; W. Ward, Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S. V., 2011).

Corpus Development: During the first 3 years of the project, data were collected from human tutored sessions and from "Wizard of Oz" (remotely controlled) virtual tutor sessions. During the 4th year, an assessment was conducted in which data were collected from students using the virtual tutor without assistance. All dialogs sessions were recorded and transcribed. A total of 427 Human tutored sessions, 1,156 WoZ sessions and 988 assessment sessions were collected.

Analysis of potential: Year 4 of the study was devoted to summative evaluation of the MyST-SDS system. A total of 219 students in 3rd, 4th, and 5th grades in Boulder Valley School District received tutoring using MyST or from human tutors. During the 2010-2011 school year we evaluated the MyST program by comparing learning gains of students who received one-on-one tutoring sessions with the virtual tutor Marni (MyST) or with human tutors in small groups. Students were randomly assigned within classrooms to the tutoring condition (virtual or human), and these groups were compared with students from intact control classrooms. The control group had significantly less residual gains compared to treatment groups. Direct comparisons of residual gain for MyST vs. Human Tutored showed no significant differences between the two treatment groups. Post-hoc tests showed no significant differences between MyST and human tutored groups; significant differences were found between MyST and the control group ($d = .53$), and human tutored students and the control group ($d = .68$).

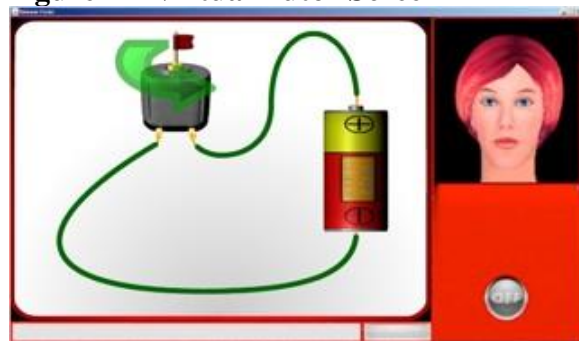
Demonstrating feasibility: The MyST tutoring treatment group in the assessment study represents the proposed intervention procedure and was implemented in 3rd, 4th and 5th grade classrooms in Boulder Valley elementary schools. In addition to the quantitative results on learning gains, we also learned that tutoring with either a virtual or human tutor engaged and motivated students, and made them more excited about science. A written survey was given to the students who participated in the 2010-2011 assessment. The survey included questions that asked for ratings of student experience and impressions of the program and its usability. (Histograms of student responses are shown in Figure 8 in the Summative Evaluation section below.) In general, students had positive experiences and impressions about the program. In general, students had positive experiences and impressions about the program. Teachers also had positive things to say about MyST and its benefits to their students. A teacher survey was administered to all participating teachers after their students completed tutoring. The survey asked teachers about the perceived impact of using Marni for student learning and engagement, impacts on instruction and scheduling, willingness to potentially adopt Marni as part of classroom instruction, and overall favorability toward participating in the research project. Some results of the survey are shown in Figure 9. 100% of responding teachers said that they felt it had a positive impact on their students, they would be interested in the program if it were available and they would recommend it to other teachers. 93% said that they would like to participate in the project again. 74% of the teachers indicated that they would like to have all of their students use the system (not just struggling students). They commented that students who used the system were more enthused about and engaged in classroom activities, and that their participation in science investigations and classroom discussions benefitted students who did not use the system.

Fifteen Conference and Workshop publications and two journal articles have resulted from the project thus far.

The Intervention – MyST-SDS

The primary goal of this project was to develop an intelligent tutoring system, My Science Tutor (MyST), intended to improve science learning by 3rd, 4th and 5th grade children through natural spoken dialogs with Marni, a virtual science tutor. MyST features automatic speech recognition, character animation, robust semantic parsing, dialog modeling and language and speech generation to support conversations with Marni, as well as the integration of multimedia content into the dialogs. Figure 1

Figure 1 – Virtual Tutor Screen



displays a screen shot of the virtual tutor Marni asking questions about media displayed in a tutorial dialog. MyST is intended to help struggling students learn the science concepts encountered in classroom science instruction. Each 15 to 20 minute MyST tutorial functions as an independent learning activity that provides the scaffolding required to stimulate students to reason and talk about science during spoken dialogs with Marni.

Marni, a lifelike 3-D computer character that is “on screen” at all times. Marni produces natural visual speech synchronized with a recorded human voice. Because Marni’s voice was recorded

by an expert science tutor, who produced prompts appropriate to the dialog context, students tended to perceive her as a sensitive and effective tutor. While talking and listening, Marni produces graceful head and face movements, including non-verbal cues like eyebrow raises and eye blinks.

In general, Marni asks students open-ended questions related to illustrations or animations displayed on the computer screen. We call these conversations with Marni *multimedia dialogs*, since students simultaneously listen to and think about Marni's questions while viewing illustrations and animations or interacting with a simulation. The system processes students' speech to assess their understanding of the science under discussion, and produces additional actions (e.g., a subsequent question that may be accompanied by a new illustration) designed to stimulate reasoning that can lead to accurate explanations. The goal of these *multimedia dialogs* is to help students construct and generate explanations that express their ideas. The dialogs are designed so that, over the course of the conversation, students reflect on their explanations and refine their ideas in relation to the media they are viewing or interacting with, leading to a deeper understanding of the science they are discussing.

MyST dialogs are linked to the activities, observations and outcomes of classroom science investigations conducted by students in the kit-based Full Option Science System (FOSS, 2007). In addition to the science kits that support an average of sixteen 30 to 60 minute investigations in each module (i.e., a specific area of science), the program includes valid and reliable standardized Assessments of Science Knowledge (ASK) administered to each student before and after each module. In our study, we developed 16 different tutorial dialog sessions, lasting about 20 minutes each, for four different FOSS modules: Magnetism and Electricity, Variables, Measurement, and Water. Thus, a total of 64 different tutorials were developed to help children think about and explain science concepts encountered during classroom activities. During these conversations, students learned to reflect on and reason about the science they learned in their hands-on science investigations and associated classroom activities.

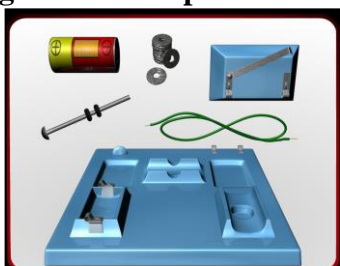
Questioning the Author: The design of spoken dialogs in MyST is based on a proven approach to classroom discussions called "Questioning the Author", or QtA, developed by Isabel Beck and Margaret McKeown (I. Beck, McKeown, Sandora, Kucan, & Worthy, 1996; McKeown & Beck, 1999; McKeown, Beck, Hamilton, & Kucan, 1999). QtA is a mature, effective, and scientifically based program used by hundreds of teachers across the U.S. It is designed to improve comprehension of narrative or expository texts that are discussed as they are read aloud in the classroom. Questioning the Author is a deceptively simple approach, its focus is to have students grapple with, and reflect upon, what an author is trying to say in order to build a representation from it. Because the dialog modeling used in QtA is well understood, can be taught to others (Beck & McKeown, 2006), and has been demonstrated to be effective in improving comprehension of informational texts. We decided to incorporate principles of QtA into the dialog strategy used in MyST. Tutors in our research study, all former science teachers, were trained in the QtA approach by one of its inventors, Dr. Margaret McKeown. Following an initial workshop in which the project tutors learned about, discussed and practiced QtA dialogs, Dr. McKeown reviewed transcriptions of tutoring sessions and provided constructive feedback to the project tutors throughout the development phase of the project. The tutorial dialogs in the final MyST system evolved from an iterative process of testing and refining these QtA-based multimedia dialogs.

Multimedia presentations play a central role in directing and focusing the dialog. Students are able to review, recall, revisit and revise their ideas about the investigation by viewing illustrations and interacting with simulations while producing and evaluating the accuracy of their self-explanations during their conversations with Marni. MyST dialogs typically incorporate three types of media: 1) static illustrations, 2) simple animations and 3) interactive investigations. Although they may overlap in the content presented, each media type plays a unique role in science learning in MyST dialogs.

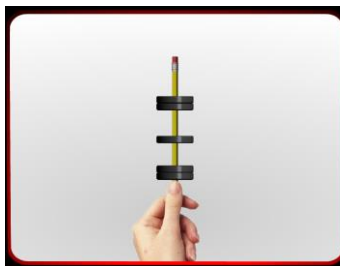
Types and Uses of Media in MyST

Static Illustrations: Static Illustrations are inanimate Flash drawings, and are a good way to initiate discussions about topics. They provide a visual frame of reference that helps focus the student’s attention and the subsequent discussion on the content of the illustration. For example, each of the illustrations in Figure 2 can be presented with questions like: “So, what’s going on here?” or “What’s this all about?”

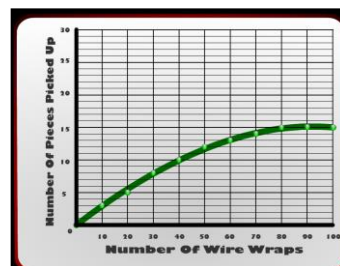
Figure 2: Example Static Illustrations



a) Objects used in class room experiments



b) Attraction and Repulsion of Doughnut Magnets



c) Graphing and Predicting Experimental Outcomes

In discussing a concept, Marni begins with indirect, open-ended questions about the illustration and then moves to increasingly more directed questions contingent on student responses. A series of questions for the first illustration in Figure 2 might be:

- *What are these things all about?*
- *You mentioned making a circuit. Tell me more about a circuit.*
- *Great thinking! What’s important about the components in a circuit?*
- *You said something interesting about components in a circuit having contact points. What are contact points all about?*

A visual like the graph could be very helpful when working with a student that grasps what they are looking at, but not how to interpret it. A QtA inspired sequence might be:

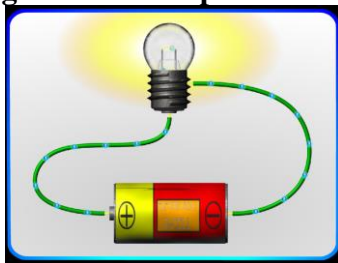
- T: What do you think this is about?
- S: I think it’s a graph of something.
- T: Yes, it’s a graph. Tell me more about the graph.
- S: Umm, I’m not really sure. It has something to do with washers picked up and wraps on an electromagnet, but I can’t tell any more than that.

- T: Great, this is a graph about the number of washers an electromagnet can pick up and how many wraps of wire it has. What happens to the number of washers picked up when the number of wraps changes?
- S: Hmm, I think it, well, I think it doesn't change? I guess I don't really know.
- T: Okay, one good way to tackle a graph is to look at the data points on the graph. Here the data points are the green dots. What do you think the first data point, all the way to the left, is telling us?

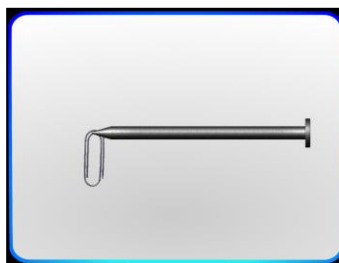
At any point that the student expresses a grasp of what a graph is, the tutor moves on to the next point.

Simple Animations: Simple Animations are non-interactive Flash animations, and can provide additional information to help students visualize concepts that can be difficult to capture in Illustrations. Figure 3 describes several simple animations, such as the flow of electricity in a circuit and the creation of a temporary magnet. In Figure 3a, the direction of the flow of electricity is represented by blue dots moving through the wires and bulb and back to the D-cell. The animations enable questions to elicit explanations about what is being shown.

Figure 3 – Example Animations



a) Electricity flowing from negative to positive terminals



b) Nail attracts clip after being rubbed by magnet

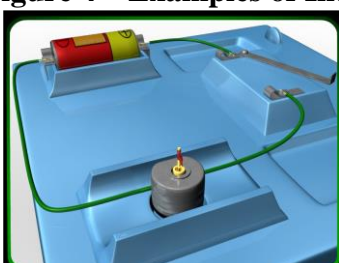


c) Magnets attract through a table

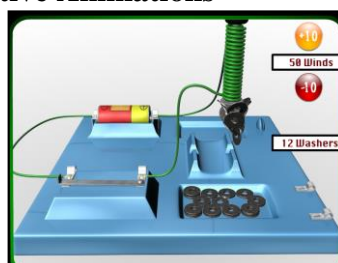
Interactive Animations: Interactive Animations allow students to interact directly with the Flash animation using a mouse. For example, clicking on the switch in a circuit will open or close the circuit, resulting in a motor running or stopping (Figure 4a), or an electromagnet picking up or dropping iron objects (Figure 4b). Interactive animations can be used to present relatively simple concepts (e.g., a switch), or to provide students with the opportunity to conduct complete virtual science investigations and graph the results. As students are interacting with a simulation, the tutor can say things like: “What could you do to ...?” “What happens if you ...?”

Each tutorial session in MyST is designed to cover a few main points (2-4) in a 15 to 20-minute session with a student. During the session, Marni attempts to elicit responses from students that show their understanding of a specific set of points, or more specifically, to entail a set of propositions. Marni attempts to elicit the points by encouraging self-expression from the

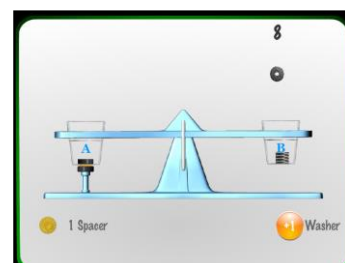
Figure 4 – Examples of Interactive Animations



a) Open and Close a Motor Circuit



b) Electromagnet simulation with changing variables



c) Breaking the Force Simulation

student. The tutorial dialog is designed to get students to articulate their ideas about concepts and be able to explain processes underlying their thinking. The strategies used in MyST to get students to share what they know are heavily influenced by QtA. Two QtA strategies that are employed by MyST are *marking* and *revoicing*. These two techniques require the ability to identify the student's dialog content (referred to as marking it) followed by repeating (revoicing) the question back to the student using similar phrasing (e.g., *You mentioned that electricity flows in a closed path. What else can you tell me about how electricity flows?*) Initially, students are prompted to consider a concept in terms of their recent experiences in class. The interactions for a concept typically begin with open-ended questions about the concept. Further sequences are written in such a way that they proceed from more general open-ended questions, "*What's this all about?*" to more directed open-ended questions, "*Tell me more about the flow of electricity in the circuit.*"

Student Interface

An example of the student's screen is shown in Figure 1 above. The student's computer shows a full screen window that contains the virtual tutor Marni, a display area for presenting media, and a display button that indicates the listening status of the system. The agent's lips and facial movements are synchronized with her speech, which may be played back from a recording or generated by a speech synthesizer (during Wizard of Oz studies only, described below). As noted, some media are interactive and the student is able to use the mouse to control elements of the display. When the student is not speaking, the listening status icon says "OFF" and is dimmed. MyST uses what is known as a "Push-and-Hold" paradigm, where the student holds down the space bar while speaking. When the space bar is released, the Listening Status indicator returns to "OFF" and the system responds to the student utterance. In interviews with students following the tutoring sessions, all students reported that they found holding down the space bar was easy to do. This procedure encouraged students to spend time thinking about their spoken responses (while Marni waited "patiently" in a state of idle animation, with natural head movements and eye blinks) before responding.

System Operation

The tutor takes a series of actions and then waits for input from the student. A typical sequence of actions would be to introduce a Flash animation ("*Let's look at this.*"), display the animation, and then ask a question ("*What's going on there?*"). Depending on the nature of the question and the media, the student may interact with content in the display area, watch a movie, or make passive observations. When ready to speak, the student holds down the space bar. As the student speaks, the audio data is sent to the speech recognition system. When the space bar is released, the single best scoring word string is sent to the parser, which returns a set of semantic parses. The set of parses is sent to the dialog manager which selects a single best parse given the current context, integrates the new information into the context and generates an action sequence given the new context. The actions are executed and the system again waits for a student response.

Each tutorial dialog is oriented around a set of key concepts that the student needs to master to understand and explain the science through the FOSS activities in the classroom. The tutoring sessions help students achieve a deeper understanding of the science as they learn how to engage in scientific discourse with Marni and construct accurate answers. The development process benefits greatly from the material provided by FOSS, which describes the key concepts in the

investigations and identifies the learning objectives. The key points for a dialog are specified as propositions that are realized as semantic frames. The tutor attempts to elicit speech from the student that entails the target propositions. Following QtA guidelines, a segment begins with an open-ended question that asks the student to relay the major ideas presented in a science investigation. Follow-up queries and media presentations are designed to 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.

Throughout a dialog, the system analyzes utterances produced by the student and maintains a context that represents which points have been correctly addressed by the student, which have been incorrectly expressed, and which have not been addressed. In analyzing a student's answer, MyST checks whether the correct entities are filling the correct semantic roles, and generates questions about the missing or erroneous elements to attempt to elicit new information about them. In the tradition of other systems using children's speech (Mostow & Aist, 1999; Mostow & Aist, 2001), MyST does not use the information extracted from students' responses to grade students, and the system never tells the student that a response is wrong. This is a good strategy for ASR-based systems because the recognizer can make mistakes. After each spoken response produced by a student, the system decides whether the current point should be discussed further, whether to present an illustration, animation or investigation accompanied by a prompt, or to move on to another point. In sessions where the system is able to accurately recognize and parse student responses, it is able to adapt the tutorial dialog to the individual student. It may move on as soon a student expresses an understanding of a point, or delve more deeply into a discussion of concepts that are not correctly expressed by the student. It may present more background material if the student doesn't seem to grasp the basic elements under discussion. If the system is unable to elicit student responses that fill any of the semantic roles related to the science concepts in a dialog, it will end up using a default tutorial presentation.

In cases where the system understands the student, it is also able to apply *marking* and other techniques that use information from the student's response to generate a follow-on question. These dialog techniques are designed to assure the student that Marni is listening to and understands what the student is saying. Marni does not simply recognize and parrot back keywords spoken by the students. It represents the events and entities in the student's response, and it also represents the relations expressed between them, and communicates this understanding back to the student. The extracted representation is compared to the desired propositions to decide what action to take next.

Using spoken responses in this way provides a robust system interaction. False Negative errors by the system, in which the system misses correct information provided by the student, account for the bulk of concept errors. In this case, the system simply continues to talk about the same point in a different way rather than moving on. False Accept errors, where the system fills in an element because of a recognition error, are very rare in MyST. When they do occur, the system may move on from a point before it is sufficiently covered. Recapitulations by the system or errors by the student in later frames often catch many of these. Thus, dialogs are designed to use speech understanding to increase efficiency and naturalness of the interaction while minimizing the impact of system errors.

Stages of Tutorial Development in MyST

MyST Development Sequence

Data were collected in three basic conditions:

1. Human Tutor – The first stage of development consisted of interactions with a human tutor. Most of these interactions were recorded and transcribed. During these interactions, a human tutor trained in QtA and the learning goals of the FOSS module conducts a tutorial with a student. Student speech is recorded and transcribed. This initial phase of develop was used to identify “sticking points” during tutorial dialogs. Subsequent analyses of these dialogs led to development of illustrations, and subsequently animations and interactive simulations that were used in subsequent dialogs. When these became available, the tutors used laptops to present media during their tutoring sessions. The dialog moves and media were thus designed and refined through this iterative process of testing and refining dialogs strategies and media. The data collected in the human tutoring sessions are used to create an initial WOZ system.
2. “Wizard of Oz” – The WOZ interface is used to interact with the student as described below. In WOZ interactions, students interact with Marni, while human tutors monitor and are able to take control of the system to produce Marni’s prompts and present media. The WOZ system is used to gather data that more closely models the desired interactions between Marni and students in the final system. These data are then used to tune the system for fully automatic operation. All interactions including student speech are saved to a time-stamped log file. The student speech is transcribed and the transcripts are automatically integrated into the log file for the session.
3. Stand-alone Virtual Tutor – Students interact with the MyST system without a “wizard” being connected. This is the procedure used in the assessment of the MyST system in schools.

Human Tutoring

The tutorial development process began with collection and annotation of dialogs between human tutors and students. These data were used: a) to train a speech recognizer to recognize the words that students produce during tutoring sessions; b) to develop natural language processing system to interpret spoken utterances; and c) to develop dialog models to interpret students’ utterances in the context of the ongoing conversation to produce responses by the virtual tutor consistent with learning objectives incorporated into the dialog model.

BLT hired an expert team of project tutors, each of whom was either a former science teacher or a science graduate student at the University of Colorado specializing in science education. Eleven tutors were hired and trained, of which 9 are still with the follow-on IES project (which includes a total of 35 tutors trained in QtA). All project staff participated in initial meetings and training sessions. These included: (a) a kickoff meeting in September 2007 with presentations by senior project personnel on each key component of the project (e.g., project overview, the FOSS science program, Questioning the Author, the process for developing dialogs, the stages of developing, testing and refining the intelligent tutoring system, and assessing outcomes); (b) a two day workshop by Margaret Mckeown explaining the Questioning the Author approach to classroom instruction and how to adapt the approach to individualized tutoring; and (c) two one-

day training sessions by Kelly Armitage on classroom instruction using FOSS science investigations for Magnetism and Electricity and for Measurement.

In order to create natural and effective interactions between Marni and the student, it is necessary to design dialogs that: 1) engage students in conversations that provide the system 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. This is the process of scaffolding learning.

A major challenge of the MyST project was how to design the spoken dialogs and media in a principled way to optimize engagement and learning. To meet this challenge, we developed an iterative approach to dialog design, informed by theory and research on learning, tutoring, and multimedia learning, in which dialogs were designed and refined through a series of design-test-refine cycles. Tutorial development followed an iterative procedure consisting of:

- Using FOSS teacher guides as a guide, project tutors develop learning objectives and supplementary materials for an investigation.
- Project tutors go into the schools and tutor students using the materials developed. The student's speech is recorded on a laptop computer and the entire session is videotaped on a DVD.
- The entire tutor group reviewed the session tapes, critiqued the presentations, and offered suggestions for improvement. A subset of the sessions was sent to Dr. McKeown who reviewed them and annotated session transcripts with comments. The tutorial presentations were revised based on the collective feedback.
- Sessions were reviewed to determine instances of misunderstandings and "sticking points" shared by several students that would benefit from the introduction of illustrations, pictures and animations that could be used to "ground" the dialogs. Sets of animations were designed and refined by the Boulder team in collaboration with Kathy Long at Lawrence Hall of Science.
- Once the tutorial content is judged to be ready, Wizard of Oz sessions are conducted, in which students interacted with Marni independently, while remote human tutors (the Wizards) monitored the session and could take control of the system when needed. The system keeps a log of each session with time-stamped entries for all events. The system logs as well as tutor comments are analyzed to find problems and suggest refinements.

Wizard of Oz (WOZ) system and data collection

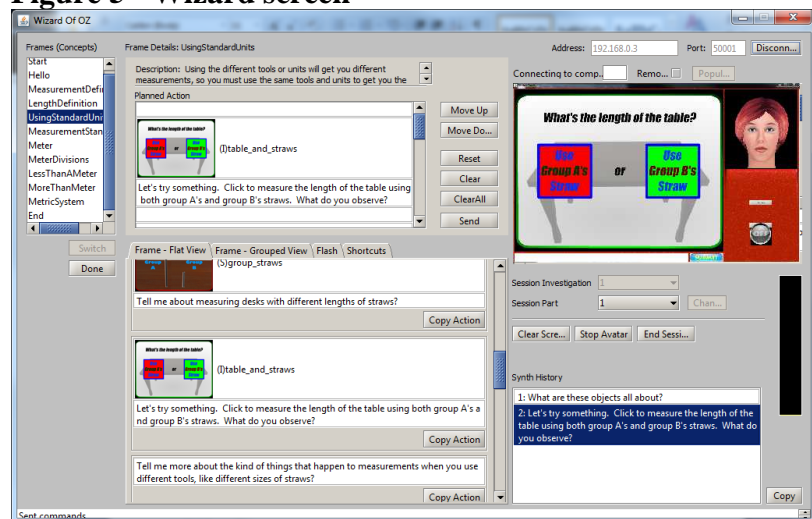
Our development strategy was to model spoken dialogs from *human tutoring sessions* of the type we would like to emulate. In order to gather and model data from effective multimedia dialogs of the sort we would like to create, we developed an interface to MyST that allows a human tutor to be inserted into the interaction loop. In this mode, the student interacts with Marni, while the human tutor can monitor the student's interaction with the system and alter system behavior

when desired. This type of data collection system is often referred to as a “Wizard of Oz” system (WOZ). The WOZ gives a remote human tutor control over the virtual tutor system. At each point in a dialog when the system is about to take an action (e.g. have Marni talk; present a new illustration) the action is first shown to the human wizard who may accept or change the action. All of the WOZ data was collected in sessions that were monitored by project tutors, who served as the “Wizards”. The data from WOZ sessions was then used to improve system coverage of concepts and to gain insights into MyST dialog behaviors based on intervention by the Wizards. During the second and third years of the project, students independently interacted with MyST in their schools, while Wizards (either at some other location at the school or at Boulder Language Technologies’ office) monitored the tutoring sessions remotely.

The WOZ interface is a pluggable MyST component that supports both independent use by a student and the ability of a human wizard to connect to any given session. If the Wizard is not connected, MyST sends the output straight to the user. If the Wizard connects to the session, MyST automatically sends actions to the Wizard for approval or revision. If the Wizard disconnects from the session, the system switches automatically to independent mode. Over the course of the data collection, we observed the expected pattern that Wizards intervene less and less as the tutorial matures during the development process. For new tutorials, Wizards intervene on an average of about 33% of the turns. This number reduces quickly to about 20%. Less than 1% of the wizard interventions involve changing the basic concepts. This implies that in almost all cases, the correct concept was being discussed by the system, but the Wizard wanted to change the specific wording in some way.

Since the WOZ interface connects to the virtual tutor over the internet, the Wizard can be at a remote site. The Wizard can see everything on the student’s computer, and hear what the student is saying, and controls system behavior using the MyST WOZ interface. Figure 5 shows the layout of the Wizard display, which contains:

Figure 5 - Wizard screen



- A screenshot of the student’s screen
- The action Marni is about to take
- The frame in focus, including all action sequences associated with elements of the frame
- A list of all frames for the session
- A set of command buttons
 - stop agent

- clear screen
- end session
- An input history list that can be recalled, to see what has been done and to allow cutting and pasting new responses.

When Marni suggests an action, it is displayed in the top-center screen. Wizards can choose to:

- Accept the proposed action
- Select a new action from the current frame
- Switch to a new frame and have the system generate a new proposed action
- Generate a new response manually by selecting system content and typing in strings for the agent to speak.

The system keeps a log of time-stamped events occurring during the session, including any Wizard-generated actions. The log records whether the Wizard accepts each proposed system action, or how they changed it. Throughout the project, we used WOZ collected data to train speech recognition acoustic and language models, and to develop grammars for parsing. An analysis of log-files from WOZ sessions gives insight into problems with tutorials and can lead to development of additional multi-media resources or modifications to cause the system to behave more like the Wizards. Analysis of the logs is used to assess the quality of the system decisions. The dialog design process incorporates analysis of transcripts of dialogs to identify the main “sticking points” that are observed by project tutors. Transcriptions have been sent to Dr. Mckeown, who reviews the dialogs and provides feedback and suggestions. Tutors review the transcripts to gain insights into strengths and weaknesses of the dialogs. The most common outcome of this process is the design of several types of media that serve to focus the conversation. Analysis of transcripts demonstrates that invoking media provides great benefit to students who have difficulty expressing their knowledge of science.

System Development

The final phase of development focused on developing and testing the fully automatic MyST-SLS system that students would use independently during the summative evaluation. MyST incorporates a number of technologies including speech recognition, dialog management, character animation, speech output, and presentation of flash applications. The system components that had already been developed were extended to be able to present flash animations concurrently with having conversational interactions with the student. For example, the system can be presenting an animation illustrating a concept; while the student is explaining what is going on in the animation, the speech recognition and dialog management system are decoding what is being said by the student. An entirely new dialog manager was developed that allows a much more conversational interaction about concepts by representing target propositions and comparing what users say to them in order to generate follow-up actions by the system.

Data Collection and Corpus Development

One significant product of the MyST project is the development of a corpus of elementary school students interacting with the virtual tutor. The Speech Recognition, Semantic Parsing and Dialog Management components of the system all require user data to develop. The corpus can be used

to train and evaluate children's speech recognition and spoken dialog algorithms. Audio recordings are transcribed and used to train acoustic models and language models for the speech recognizer. The transcripts are also used to develop grammars for the semantic parser.

The corpus can also support other research efforts such as analyzing the characteristics of children's speech and determining features that are associated with learning gains. At the completion of the project, the corpus, which will contain over 150 hours of children's speech during tutorial dialogs, will be made available to the research community.

All data were collected from sessions at elementary schools in the Boulder Valley School District (BVSD). BVSD is a 27,000-student school district with 34 elementary schools. There is great student diversity across schools, which vary from low to high performing on state science tests. We administered tutorial dialogs to students in both high performing and low performing schools in order to gauge the potential benefits to a broad range of students.

Speech Files

The speech data are stored in files by student turns, i.e. whatever is said from the time the student pressed the space bar to talk until the bar is released. The speech is sampled at 16 KHz, as is typical with microphone speech. The subjects are wearing Sennheiser headsets with noise canceling microphones. The speech data are professionally transcribed at the word level. Disfluencies (false starts, truncated words, filled pauses, etc) are also marked in the transcriptions.

Log files

Each MyST dialog session produces a log file that contains time-stamped entries for the events that occurred during the dialog. At each point that the student speaks, an entry is written into the log that gives the filename for the associated recorded speech file. The speech recognition output is logged. Manual transcription of the speech files is performed off-line and is introduced into the log file later. Some additional pieces of information stored in the log file are: extracted frame elements, current context, frame name and frame element or rule that is generating the system response, the number of times this frame element or rule has been used, and the action sequence generated for the response. Following manual transcription of students' speech during dialogs, scripts were written to process the log files to gain insights into the way in which students interacted with Marni, how different system behaviors affected learning, and how the human language technologies performed.

Concept Annotation

The transcript data are annotated to mark the concepts used by the semantic parser. Human annotators highlight word strings in the transcripts and assign the appropriate concept tags. The concept annotations are hierarchical, for example *from the positive end* would be a :DirFlow:::Origin:::Terminal: concept where the substring *positive end* refers to a :Terminal: of a battery. This process is essentially finding paraphrases of the ways concepts are referred to. These annotations are used to expand the coverage of the grammar patterns for the parser, to evaluate coverage of the parser, and to provide "gold standard" input for testing other components of the system.

MyST System Component Evaluations

The collected data were partitioned by speaker into training, development, and evaluation sets. Data from any individual student was in only one of the sets. The training set was used to train acoustic models and language models for the speech recognizer and to train grammar patterns for the parser. The development set was used to optimize parameter values such as language model weights. The evaluation set was used for component level evaluation of the ASR and parsing components.

Automatic Speech Recognition Performance

The recognizer is trained and parameterized using the training and development data and run on the evaluation set using a language model (trained on all training data), that has a perplexity of 63 for the evaluation set. The vocabulary size was 6,235 words. The Word Error

Table 1 - Results for Speech Recognition

	Baseline		+VTLN		+VTLN +MLLR	
	WER(%)	CA	WER(%)	CA	WER(%)	CA
ME	29.8	.85/.89	28.1	.87/.91	26.1	.87/.91
MS	29.6	.83/.87	28.6	.84/.87	26.7	.86/.89
VB	36.1	.82/.89	34.3	.80/.87	31.9	.82/.90
Tot	30.9	.84/.89	29.5	.85/.89	27.4	.86/.90

Rate (WER) for the recognizer on the Evaluation set is shown in Table 1 in the *Baseline* column. The Out of Vocabulary word rate was very low for all modules, ranging from 0.6% for Magnetism and Electricity to 0.7% for Variables. There were a total of 65,496 words in the evaluation set.

The WER for the pooled data (Tot) was 30.9%. These baseline results were obtained using speaker-independent acoustic models, but not adapted to the current user. A number of speaker adaptation techniques are commonly used in ASR systems. Two of the most effective are Maximum Likelihood Linear Regression (Leggetter & Woodland, 1995) and Vocal Tract Length Normalization (Lee & Rose, 1998). Vocal Tract Length Normalization (VTLN) is motivated by the fact that different speakers have vocal tracts of different length, which results in a variation of the formant frequencies. VTLN compensates for this variability by applying a warping factor to the speech spectrum in the frequency domain. For each speaker, a first pass of the decoder was run to generate a hypothesis word string. A warping factor was then computed for the speaker to maximize the likelihood of the features extracted from the speech given the hypothesis. This warping factor is then used to produce a final hypothesis in a second decoding pass. The application of VTLN reduced the WER from 30.9% to 29.5%. MLLR works in the acoustic model space, rather than feature space like VTLN, and consists of applying a set of transforms to the Gaussian means and co-variances of the speaker independent acoustic models to better match the speech characteristics of the target speaker. Transforms are estimated so that, when applied to the parameters of the acoustic models, the likelihood of the speaker data is maximized with respect to the hypothesized sequence of words. Speaker data are then re-decoded after applying the transforms. The number of transforms is determined dynamically based on the adaptation data available. Adding MLLR adaptation reduced the error rate further to 27.4%.

For the numbers listed above, the adaptation techniques were applied in a batch unsupervised mode using all of the data for the particular speaker. In a live application, for new users, warping factors and transforms would need to be computed incrementally as more data come in, or after a certain minimum amount of speech data were available. The benefits of adaptation would initially be small and should improve rapidly as more speech data become available. In this intervention (MyST), it is anticipated that an individual student will use the system repeatedly over a period of time. A single FOSS Module will have 16 tutorial sessions associated with it, each lasting about 20 min. The cumulative data from each user will be used to pre-compute warp factors and transforms that are stored and loaded when the user logs in. On average, first time users will initially experience system performance similar to that in the Baseline column in Table 1, WER of around 31%. The system will incrementally adapt as more data from the user are available over sessions. Since the batch unsupervised adaptation described above not only adapts to the speaker, but also to the test data, performance in live use would not be expected to fully reach the same level of performance.

Concept Accuracy

The behavior of the virtual tutor is more dependent on Concept Accuracy than on Word Error Rate. One way to measure the effect of recognition errors on the system is to look at the accuracy of extraction of frame elements. Grammars are created for each investigation using the training data. The investigations have an average of 8 frames with an average of 5 frame elements per frame, thus there are about 40 frame element classes on average in an investigation. Reference parses were created for each hand transcribed utterance by parsing the transcripts, which represent word input with no ASR errors. The speech recognizer output for the utterances was also parsed and Recall and Precision of frame elements were calculated compared to the reference parses. Recall is the percentage of the reference elements that were correctly extracted from the recognizer output. Precision is the percentage of the elements extracted from the recognizer output that were correct. The results for Concept Accuracy are shown in the columns labeled CA in Table 2. The first number in the accuracy is Recall and the second number is Precision. Using a global LM, the baseline system had a WER of 30.9% with an overall Recall of .84 and Precision of .89. With batch unsupervised speaker adaptation, a WER of 27.4% with a Recall of .86 and a Precision of .90 were achieved.

Summative Evaluation of MyST-SLS

During the 2010-2011 school year we evaluated the MyST-SLS program by comparing learning gains of students who received tutoring sessions soon after classroom science investigations with either the virtual tutor Marni (MyST) or with human tutors in small groups. Students were randomly assigned within classrooms to the tutoring condition (Virtual or Human), and these groups were compared with students from intact control classrooms. Students completed one of four FOSS modules-- *Variables, Magnetism & Electricity, Measurement, and Water*. All students received similar classroom instruction.

The hypotheses for the study were: 1) students in MyST and human-tutored groups would have roughly similar gains from pre to post test, 2) tutored students would have significantly greater gains than students in the control (nontreatment) conditions. The complete report on the assessment is included in Section C, and a brief summary is presented here.

The FOSS Assessing Science Knowledge (ASK) instruments were used to measure learning gains for each of the four modules in the study. The ASK assessments consist of identical pre and post versions with open-ended, short answer, multiple choice and graphing items administered before the beginning of the FOSS lessons, and immediately after classroom instruction and tutoring ended. Pairs of raters from Boulder Language Technology scored all assessments from tutored students, and a subset of students from control students. All scoring was blind to tutoring group. Inter-rater reliabilities for two raters were high (counting only the open-ended items) with intra-class correlation coefficients ranging from 0.89 to 0.98. Internal reliabilities were lower, ranging from 0.60 to 0.89 for both pre and post versions of the assessments. Scores used for outcome analysis were the averages across both raters.

Research was conducted at schools with students from a large range of socioeconomic and ethnic backgrounds. Eighty-three (83) students received MyST tutoring, 69 were human tutored (both in 12 classrooms) and 1015 students in 50 classrooms in 20 schools received only classroom instruction and no tutoring. Sixty-two (62) classrooms were included in the analysis. To make comparisons, outcome scores were converted to *Residual Gain Scores*, which compared groups on the average differences between their observed and expected scores. Additionally, residual gain scores were estimated and evaluated assuming and not assuming equal variances. The difference in *t*-value was only 0.01, and did not affect the associated significance levels.

Direct comparisons of residual gain for the treatment groups (MyST and Human Tutored) showed no significant differences between the two treatment groups with $t = -1.14$, $df = 150$, $p = 0.25$. This supports Hypothesis 1, that learning gains from using MyST would be roughly similar to gains produced by human tutors. In the three-way comparison with the control group, MyST and human tutored groups had insignificantly different residual pre/post gains; the control students, on the other hand, had significantly less residual pre/post gains. A Univariate ANOVA (using scores standardized by module test) showed a main effect for tutoring condition with $F = 26.2$, $df = (2, 1164)$, $p < 0.01$. This supports Hypothesis 2, that both tutored groups would have greater gains than the control. Post-hoc tests showed no significant differences between MyST and human tutored groups; significant differences were found between MyST and the control group (d

Figure 6 – Residual Gains
Effect Size for Residual Gain Scores by Group for Pre/post FOSS-ASK

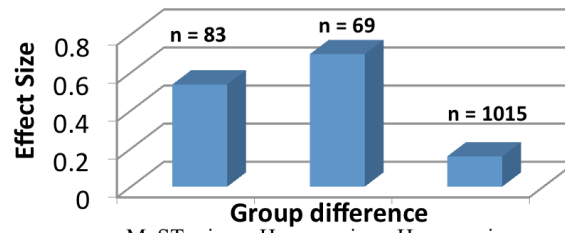
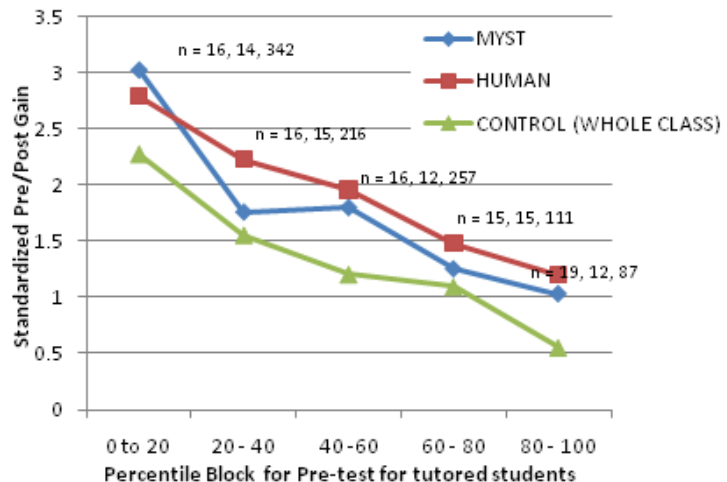


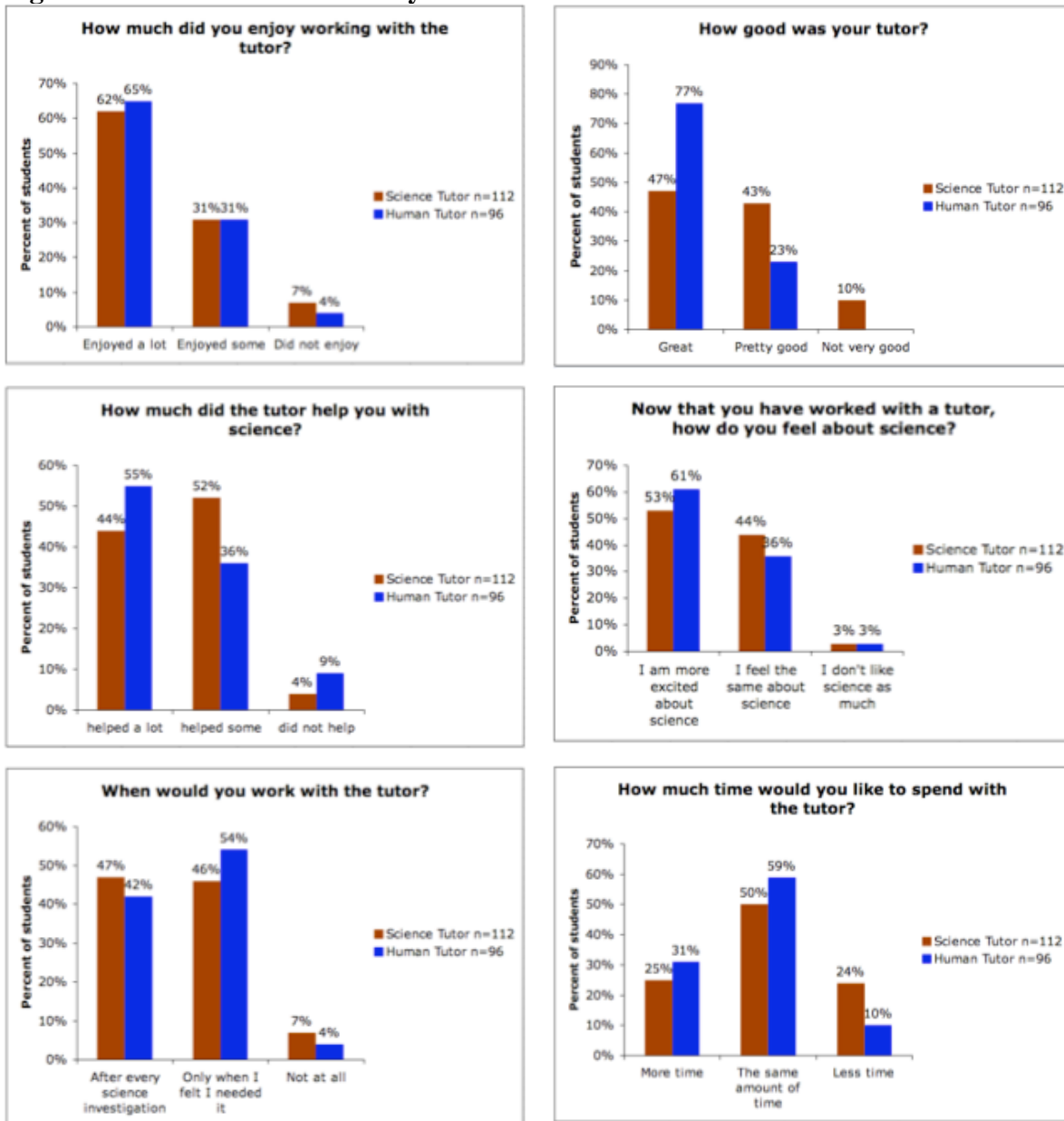
Figure 7 – Residual Gain by Pre-score
Standardized Pre/Post Gain on FOSS-ASK by Percentile Block of Pre-test



=.53), and human tutored students and the control group ($d = .68$). Differences in residual gain scores were also tested using hierarchical models with classroom used as a grouping variable. MyST students showed significantly higher scores than the controls ($t = 2.5, df = 60, p = 0.014$), as did the human-tutored group when compared with controls ($t = 3, df = 60, p < 0.01$). Differences between group means for residual gain score also varied by where students scored on the pre-test. Figure 6 shows that struggling students benefited most from MyST and human tutoring. That is, MyST and human tutoring had the greatest effect on the lowest performing students based on their pretest scores, and the least effect on students with the highest pretest scores, with decreasing benefit for both tutoring conditions across the five quintile groupings.

Is it Feasible to Integrate MyST-SLS into Elementary Science Curricula?

Figure 8 – Tutor vs Marni Survey Results



The MyST tutoring treatment group in the assessment study represents the proposed intervention procedure in real world educational settings. The study thus represents an initial investigation of the feasibility of integrating MyST into classroom science instruction. In our study, students left their classrooms to use the system during specified times that did not interfere with structured classroom instruction, lunch, music, physical education or playground time. Project staff went to each classroom to bring consented students to laptops that were provided for the project in spaces designated by the school. These spaces varied widely, from little used hallways, to libraries or resource rooms. Because of these constraints, the implementation of MyST into elementary school classroom science instruction probably does not speak to realities of how principals and teachers would integrate it into instruction if the program was a fully developed commercial product. Nevertheless, MyST was used by approximately half of all students in each participating treatment classroom, and was used consistently by all of these students after conducting classroom science investigations in four different areas of science. The study therefore provides some initial insights about teachers' impressions of MyST as a tool that could be integrated into classroom science instruction. (We note that the IES Goal 3 grant awarded to BLT in June 2013 is designed to replicate and demonstrate the efficacy of MyST. In the two year efficacy study, students will use MyST independently in classrooms or resource rooms without any supervision by project staff. This study is expected to answer questions about the feasibility of integrating MyST into classroom science instruction.)

A written survey was given to the students who participated in the 2010-2011 assessment. Measures were taken to avoid bias wherein students give overly positive answers to questionnaires, including: 1) written (versus oral) surveys for students were administered, 2) students were verbally assured of anonymity, 3) questionnaires were anonymous in that students did not write their names on the survey, and 4) adults from the program did not directly observe or interfere with students while they completed the survey. The survey included questions that asked for ratings of student experience and impressions of the program and its usability. Three point rating scales for survey items were keyed to each question. A typical question, such as: *How much did Marni help with science?* had responses such as: *Did not help, helped some, helped a lot.* Items were written to reflect the reading level of the students. Histograms of student responses are shown in the Figures 8. In general, students had positive experiences and impressions about the program. Across schools, 47% of students said they would like to talk with Marni after every science investigation, 62% said they enjoyed working with Marni "a lot," and 53% selected "I am more excited about science" after using the program. Only 4% felt that the tutoring did not help. One unanticipated result was that students whose parents did not originally sign the consent form allowing their child to work with Marni often asked their parents to sign the form after learning how much other students enjoyed the experience.

Teachers were asked for feedback to help assess the feasibility of using MyST as a supplement to classroom instruction, and to share their perceptions of the impact of the system on their students. A teacher survey was administered to all participating teachers directly after their students completed tutoring. Teachers were assured anonymity in their responses both verbally and in written form. The questionnaire contained 22 rating items as well as 9 open-ended questions. The survey asked teachers about the perceived impact of using Marni for student learning and engagement, impacts on instruction and scheduling, willingness to potentially adopt Marni as part of classroom instruction, and overall favorability toward participating in the

research project. Additionally, teachers answered items related to potential barriers in implementing new technology in the classroom.

The 43 different teachers whose students used either MyST-SDS or MyST-MP&D had generally positive impressions of the system, their students' experiences using it, and the systems' likely benefits to their students. In figure 9 below we combined the responses of teachers whose students used MyST-SDS and MyST-MYP&D since a) the teachers' responses were very similar across the two systems, and b) since students left the classrooms when they were tutored, and teachers did not observe students using either system, teachers' impressions were based on students' science learning and their interactions with other students after using the system.

All teachers reported that the MyST system had a positive impact on their students and that they would recommend the program to other teachers. All but one teacher said that they would like to use the program again in the future. Interestingly, teachers indicated that, if given the choice, they would have all of their students use MyST, rather than just struggling students. Teachers also commented that students who used the system were more enthused about and engaged in classroom activities and that their participation in science investigations and classroom discussions benefitted students who did not use the system. Histograms of the teachers' responses to survey questions are shown in Figure 15 below.

2. MyST-MP&D

MyST-MP&D was developed to investigate an alternative approach to tutorial dialogs, which combines multimedia presentations of science followed by question-answer dialogs. There are two main differences between MyST-SLS and MyST MP&D. First, in MyST-SLS, the overarching goal is to have students learn by constructing explanations, with the virtual tutor scaffolding learning through questions and media; explicit teaching is limited to brief summaries of key concepts are logical points during the dialogs. In MyST-MP&D, children are presented with narrated animations that explain science based on established principles of multimedia learning that optimize retention of information and transfer of knowledge to new scenarios (Mayer, 2005). The idea is that students will receive an explanation that will help them understand and visualize the science, and provide a sufficient level of understanding to reason and talk about it. The multimedia presentations are followed by a question-answer dialogs that assesses students' understanding of the science using thoughtful multiple choice questions (MCQs) with challenging answer choices, with immediate formative feedback provided following selection of answers. The session concludes with a brief spoken dialog with the virtual tutor.

Second, MyST-MP&D was designed to support both *one-on-one tutoring* and *tutoring in small groups of 3 students*. Students within classrooms were randomly assigned to one of these two conditions. All students had dialogs with Marni, in either one-on-one or small group sessions.

Sequence of MyST-MP&D Activities

Title Screen: Each MyST-MP&D session began with a title screen that presented a deep reasoning question. In all cases, the printed question was read aloud by the virtual tutor. Examples included: What do magnets stick to? What is an electrical circuit? How can we measure length (volume, mass, temperature) and get the same answer each time? The tutoring session was introduced with an authentic question; research indicates that presenting authentic questions that require students to think about the topic before instruction begins improves learning (Driscoll et al., 2003; Gholson et al., 2009; Sullins, Craig, & Graesser, 2010).

Engaging Real-life Scenario: The first multimedia presentation was a narrated animation that introduced the science. It associated the science with materials and situations likely to be familiar to most or all of the students. The Scenario was designed to help students make meaningful connections between the science and their own experiences and knowledge, to introduce and discuss scientific vocabulary and concepts, and to them make connections between the scenario and the deep reasoning question introduced on the title screen the MYST-MP&D session.

Multimedia Science Explanation: Students were presented with a multimedia presentation that explained the science. The design of these multimedia presentations is based on a substantial body of theory and research in multimedia learning. This literature informs the design of narrated animations that optimize learning and support development of rich mental models that integrate verbal and visual information (Mayer, 2005). In MYST-MP&D explanations, each narrated animation is consistent with the multimedia principle of *segmentation*. Each narrated animation sequences the presentation in terms of the underlying set of scientific concepts, with brief pauses between each segment, so concepts build on each other to support a complete and accurate explanation. For example, the concepts underlying an electric circuit include: a circuit is a

complete pathway through which electricity flows, electricity flows from the source of the electricity through the receiver and back to the source, and electricity flows in one direction only, out of the negative side of the battery and back into the positive side. Figure 10 presents an example of a multimedia explanation for measurement.

Figure 9

TITLE a


How do you measure accurately?

START

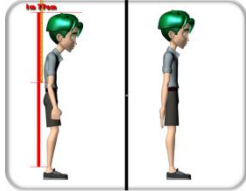
Teacher initiates CASUM by loading up title screen. Then they have kids read / write and think about the question.

When ready they begin the CASUM tutorial by clicking on start.

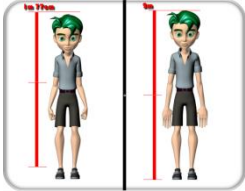
SCENARIO b




“Today I measured how tall Jack was.”



“The first time I measured him he was 1 meter 77 cm tall”



“The second time I measured him he was 2 meters tall”



“What’s going on here?”

“How do you measure accurately?”

PAUSE: Engage in conversation. Reiterate deep question and then collect students ideas and have them build off each other.

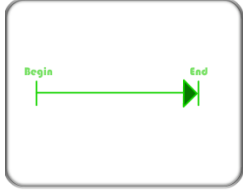
REPLAY

CONTINUE

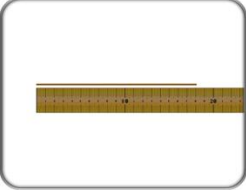
Suggested actions: Replay again, discuss what they notice going on about the quality of measurements that Jill is taking and why they are different. Then click on Continue. OR, just Continue to EXPLANATION.

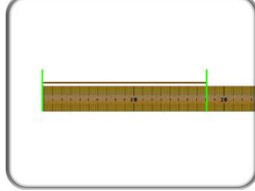
Figure 10

EXPLANATION



“When measuring length, it is important to begin and end your measurement at the right places.”



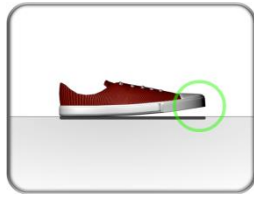


“It is also very important to make sure things are flat and lined up with your meter stick.”

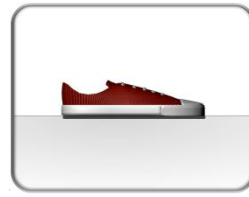
PAUSE: Teacher could pause and connect this visual to the first one so kids see that the thing you measure has a distinct beginning and end -What are the green lines all about? -How do they connect to what is important about measuring length?



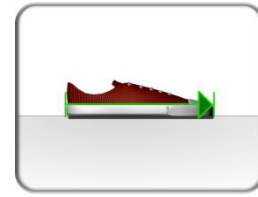
“Take this shoe for example.”



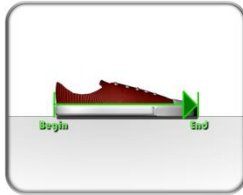
“Do you see how the tip of the shoe is lifted up a bit?”



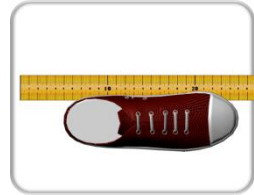
“In order to get a good measurement, we first have to make sure that the object we are measuring is as flat as possible.”



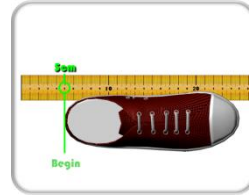
“Then we start our measurements at the back of the shoe
And measure all the way to the tip of the shoe.”



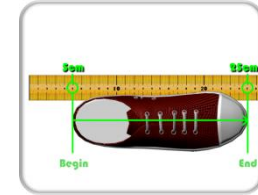
“These two spots are the beginning and end of our shoe”



“But we need our meter stick to make an actual measurement.”

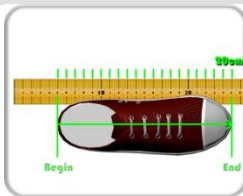


“When we look at the meter stick, we see that the back of our shoe is at the 5 cm mark...”



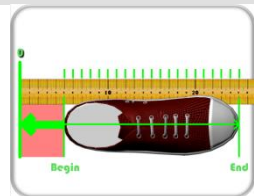
“...and the tip of our shoe is at the 25 centimeter mark. Does that mean our shoe is 25 cm long?”

PAUSE: This is a good time to pause to review what they have seen happening and how that connects to what they think are ways to measure accurately.

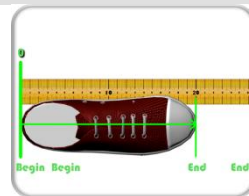


“Well, when we actually count the centimeters starting at the back of the shoe and ending at the tip of our shoe...we count twenty centimeters, not twenty-five.”

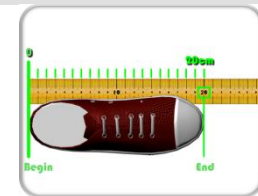
“Oh, instead of counting the units in-between, is there an easier way?”



“Sure, the easiest and best thing to do is to just move the shoe to the zero mark and start your measurements from the end of the meter stick.”



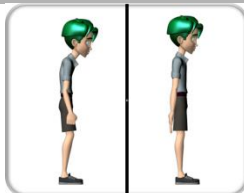
“And see what happens? Our shoe starts at the zero mark and ends at the twenty mark.”



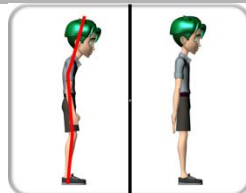
“We measured the shoe as being twenty centimeters long. That’s the same measurement that we got before. Perfect!”

Figure 11

SOLUTION



“Well Jill when you measured me you got two different measurements. Maybe how I was standing was part of the difference in those measurements.”



“The first time I wasn’t standing straight, I was hunched over.”



“And my first were not close to the wall, so I was not flat against the wall.”



“But the second time I did put my feet close to the wall And I pressed my back up...”

<p>“...so I was nice and straight all along the wall.”</p>	<p>“And remember the first time the meter stick was not down on floor by your feet. It was up by your calf above your ankle. See, I started at the wrong place.”</p>	<p>“So this time I’ll measure the first meter...”</p>	<p>“...and mark it right here.”</p>
<p>“Then I’ll move the meter stick up and line it up with the mark.”</p>	<p>And finally I can measure the last length, which is a nother full meter.”</p>	<p>“Then I add the two measurements together: one meter plus one meter is two meters! That’s the same measurement we got before when I also used goo measuring techniques”</p>	<p>“So you are two meters tall; that’s pretty tall Jack.”</p>
	<p>“Well, now that we figured it out...that was pretty easy.”</p>		

Formative Assessment (MC Question): After the multimedia presentation is completed, students were presented with an authentic question that can be answered if students have achieved a deep understanding of the science. The question was sometimes the same as the deep reasoning question that introduced the MYST-MP&D session. In some tutoring sessions, a different question was presented. Questions were often accompanied by illustrations, and required answers that demonstrated application of the science knowledge to the situation shown in the picture.

Spoken Response to the Authentic Question: Following presentation of the question, students were asked to produce a spoken answer to the question *before the four answer choices were presented*. The goal was make students think about the question, and express their understanding in words. We note that students in the small group condition were encouraged to discuss the question and to attempt to converge on an explanation.

After producing a spoken response, students were presented with the question a second time, along with the four response alternatives. Students were required to listen to the virtual tutor read each answer choice aloud and were then asked to select the best answer. All choices were presented for two reasons: a) some answer choices were correct, but were not the best (e.g., most complete) answer to the question, and b) we wanted to be sure that students in small groups listened to each question so students could discuss the answer choices. After an answer was

selected, virtual tutor provided immediate formative feedback on the choice; if an incorrect answer was selected, the tutor explained why it was incorrect, then presented the correct answer, and expanded upon why it was the correct one.

Spoken Dialogs with Marni: Each session concluded with a spoken dialog with Marni, lasting less than 5 minutes. These were truncated versions of the MyST-SLS dialogs, in which Marni asked an open-ended question designed to elicit a complete and accurate explanation of the science phenomena or systems in the multimedia presentations. If the explanation was not complete, Marni asked follow-up questions. Students in small groups were encouraged to discuss their answers before the designated speaker responded.

Quantitative Results

MyST-MP&D Summative Evaluation

Hypotheses

The two hypotheses for the study were:

- 1) *Students receiving computerized tutoring in groups will achieve learning gains similar to students receiving one-on-one tutoring.*
- 2) *Both groups receiving tutoring will gain more from pretests to posttest than students receiving no tutoring.*

We did not expect statistically significant differences learning gains between students in who received one-on-one tutoring and students who received small group tutoring; we expected both groups to benefit from tutoring, and achieve gains similar to those obtained in the 2010-2011 study for one-on-one and human tutoring.

Research Design Procedures

The assessment of the MyST-MP&D treatments was conducted from November 2011 to May 2012 in 3rd and 4th grade classrooms in the Boulder Valley School District. All students in the 2011-2012 study received in-class instruction in either the FOSS module *Magnetism and Electricity* (4th grade) or *Measurement* (3rd grade). Participating teachers followed module lesson plans and had their students conduct all science investigations. The duration of instruction using the FOSS science modules varied from one to three months during the school year.

One hundred eighty-three students in 13 classrooms at four schools participated in the study. Of the 183 students, 114 were randomly assigned to the “group” experimental condition and 69 were in the “individual” condition, with 100 students completing the FOSS *Magnetism and Electricity* module and 83 completing the *Measurement* module.

Students in the small group condition were encouraged to discuss answers to Marni’s questions. Each student sat in front of a laptop computer wearing headphones so they could look at and listen to Marni when she talked, and view and listen to the narrated presentation. In each

Table 3

Means, Standard Deviation, Pre/Post Average and Scale for FOSS-ASK tests.					
Module		Pre (raw)	Post (raw)	Pre/Post Average	Scale
Measurement	Mean	27.8	44.6	36.2	9 – 63
	SD	8.7	7.9	8.3	
M & E	Mean	21.2	29.6	25.4	7 – 39
	SD	7.4	8.8	8.1	

session, only one of the students in the group communicated with Marni; students took turns being the speaker. We note that the MyST system did not record and process discussions among students in small groups. Marni listened to and responded only to the designated speaker in each session. Project tutors observed each group session, and coded students' conversations, as discussed below.

Students in the Group condition worked in groups of three (except when a student was absent) and responded to questions about the multimedia science presentations. Typically, the group leader (the designated speaker for the session) asked the other students to confirm his or her answer, or asked others if they knew the correct answer. After discussion the group leader gave the agreed upon answer to MyST. Students in the one-on-one treatment interacted directly with MyST by answering questions verbally, or by choosing multiple choice answers.

Analysis of Learning Gains

Measures and Scores: The FOSS - ASK assessments for the two modules used in the assessment have identical pre and post versions with open-ended, short answer, multiple choice and graphing items. Tests were administered before the beginning of the FOSS lessons, and immediately after tutoring ended at the school. Students completed pre/post FOSS-ASK assessments for *Measurement* and *Magnetism & Electricity* modules before and after the classroom instruction and tutoring. Learning gains from pretest to posttest for students in the individual and small group tutoring treatment conditions were compared to learning gains of students in classrooms in the 2010-2011 MyST-SLS study who received classroom instruction for *Measurement & Magnetism & Electricity* who did not receive supplemental tutoring.

Standardization: Because module tests have different scales (see table 3), scores were standardized to a common metric. All standardization used scores from both years of the study with outliers and other spurious data removed. "Test-wise" standardization subtracted the mean of each test (over all students and pooling pre/post) from each students score. This difference was then divided by the weighted average standard deviation for both pre and post for each test. Information about each test is presented in Table 4.

Table 4

Mean, SD, N and Effect Size for tutoring groups 2011 and 2012.				
Tutor Condition	Mean	Std. Deviation	N	Effect Size
MyST-SDS Tutor (2011)	0.34	0.84	83	0.51
Human Tutor (2011)	0.47	0.73	69	0.65
Control (Whole Class) (2011)	-0.13	0.93	1015	
Group MyST-MP&D (2012)	0.43	0.72	103	0.61
One-on-one MyST-MP&D (2012)	0.45	0.72	61	0.63

Note: Comparisons for 2010-11 data incorporated Variables and Water modules

Test reliability: Pairs of raters scored all assessments from tutored students. The raters were project tutors from Boulder Language Technology who were blind to subjects' treatment conditions, and whether the assessments they scored were pretests or post-tests. Raters trained together with scoring rubrics provided by FOSS, then scored the assessments independently. All scoring was blind to tutoring group and raters did not know if scores were pre or post. Inter-rater reliabilities for two raters were high (counting only the open-ended items) with intra-class correlation coefficients ranging from .89 to .94, with averages for pre and post .91 and .94.

Internal reliabilities (Cronbach’s Alpha) were lower, ranging from $\alpha = .66$ to $\alpha = .87$ for both pre and post versions of the assessments, with averages for pre = .78 and post = .78. Internal reliability varied for each module. Scores used for outcome analysis were the averages across both raters.

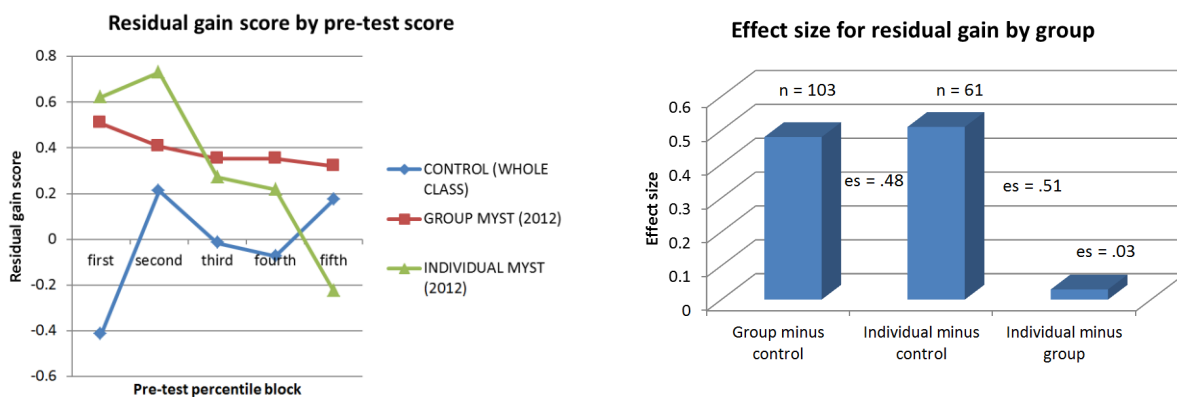
Results

When compared with the control group, effect sizes for were $d = .48$ for the Group condition and $d = .51$ for the Individual condition. These were both close to what we found for the year before for individual tutoring with MyST-SDS ($d = .51$) and for human tutors ($d = .65$). Pre/post gain differences among groups varied by FOSS module with less relative gain for experimental groups on the *Measurement* module and greater gains for the *Magnetism and Electricity* module. Lower achieving students on the pre-test in the Individual group gained relatively more than students in the control condition, but gain for these students decreased for higher performing students on the pretest. Students in the Group condition gained more than controls across the ability scale.

Gain by initial ability level.

Gain was also assessed based on ability level for the pre-score. Group comparisons divided the pre-score distribution for the tutored group in five equal parts. The resulting distribution showed higher gain for tutored groups in the lower pre-score blocks especially for the Individual group, with more uniform gain across ability for students in the Group condition.

Figure 12



Comparisons with treatment groups from 2010-2011

A Two-factor ANOVA tested if group means from both years differed significantly on residual gain score. The main effect for tutoring group for all groups (2011, 2012) was significant with $F = 16.8$, $df 4,1171$, $p < .0001^{**}$. No significant interaction was present for the treatment by module effect indicating that differences generalized to each module. Post-hoc tests showed significant differences between all tutoring groups and the control group, and no significant differences were evident among any of the four tutoring conditions. Effect sizes for MyST tutoring were higher in 2012 than 2011 although face-to-face “human” tutoring still had the largest gain.

Observations of Interactions among Students in Small Groups

We made structured observations of students interacting with MyST. Observers used a checklist that allowed observers to record the duration of student answers to questions from Marni, the types of questions asked by MyST, and the characteristics of discussions between students. The checklist was on a PDA and electronic data was imported into an Excel database. (See appendix for checklist).

We tested the reliability of the observations by having two observers watch the same students. Agreement between raters varied from 70% to 89% for type of question, and type of discussion. A sample of observations for the duration and number of student answers for a tutoring session were also checked against computer logs; differences were usually minor for number of observations (deviation of + or - 2 observations), and duration correlated highly with $r = .87$ between observation and log. Data from two observers with low agreement and anomalous ratings were removed from the dataset. Five observers observed 64 students at three schools. Two hundred eight (208) tutoring sessions were observed with 4749 observed group answers to questions and 13,430 individual records.¹

We observed how students in groups answered these questions. The group consisted of a “leader” (the student who talked with Marni using a headset microphone; and the other two members of the group (“listeners”) who contributed to answers. Students took turns across different tutoring sessions being the leader. The leader of the group was instructed to consult with the other students before answering questions. The resulting answers were divided into short *confirmational exchanges*, verses exchanges where students engaged in more interactive *discussions*. Confirmational exchanges were typically much shorter than discussions and consisted of either the group leader providing an answer and then the listeners agreeing with this answer, or a listener providing an answer, with the leader then repeating it to Marni. Discussions were usually longer in duration than confirmational exchanges, with students elaborating on each other’s answers, disagreeing with each other, or referencing previous classroom instruction. A typical discussion had multiple back-and-forth student exchanges culminating in an agreed upon answer to a question.

In some cases the group leader did not ask for input from the other students and just answered the questions. If the project tutors who observed sessions observed this occurring frequently, they reminded the group that all members should participate in discussing the answers.

Types of questions and responses

We wanted to know if specific types of questions were more likely to elicit interactive discussions. Students’ responses were analyzed for three different types of questions:

1. Initial question: This is the authentic question that students produce a spoken response to before being presented with four alternative response choices.
2. Answers to Multiple-Choice questions: Discussions students had about the four different response choices that were read aloud following the authentic question.

¹ Students were in groups of two or three; records in the databases are organized by individual observations, observation sessions, and by student.

3. Spoken Dialogs with Marni: These were students' spoken responses to open-ended questions that concluded the dialog session. These questions followed the QtA format and were designed to elicit explanations of the science displayed in the multimedia presentations.

Characteristics of interactive discussions

On average interactive discussions were 54% of all types of exchanges, and accounted for 65% of total time observed. These percentages varied widely across observations. Average discussions were 30 seconds long (versus 16 seconds for conformational exchanges).

When students did engage in interactive discussions, the majority of the time (81%) was spent elaborating on other students' comments. These comments often involved students adding new information to a leader's answers, or rewording or clarifying answers from another student. Fewer discussions involved students disagreeing with each other, which only happened in 10% of discussions; students only infrequently (3%) referred or referenced prior classroom instruction. (This is an interesting result, given that the majority of students reported on the questionnaire (see below) that they often agreed with the answer the leader gave.)

In sum, observations of students working in groups examined length and characteristics of student interactions, and linked this information with computer logs and the ASK assessment data. From these observations we found that lengthier student discussions with students elaborating on each other's' answers, disagreeing about answers or referencing classroom instruction were more frequent when a) questions were asked directly after the initial multimedia presentations, and b) During the final spoken dialog after Marni asked the first authentic question. Extended discussions were less frequent for follow-on questions during the spoken dialogs, and during consideration of answer choices to multiple-choice questions. The shorter discussions during consideration of alternative response choices to MCQs were often confirmatory discussions, in which the group quickly concurred with the answer choice selected by one of the members of the group. Based on students' responses to the questionnaire, we expect peer pressure may have been involved in these short exchanges, as students reported that they often disagreed with the answer that was given.

While students who scored higher on the pre-test tended to participate more frequently in extended student discussions, participating in discussions did not correlate with student gain from pre to post on the ASK assessment.

Links between FOSS-ASK assessments and types of responses in small groups: We also wanted to know if gain on the FOSS-ASK assessment was related to the frequency and duration of discussions. The average amount of time spent by students in interactive discussions was correlated ($r = .23$) with pre-test scores, but not with either pretest vs. posttest gain or post-test score. This result generalized for both FOSS modules. The correlation with pre-test suggests that students who score higher on the pre-test tend to also be more likely to engage in discussions.

Students and Teachers Experiences with MyST-MP&D during One-on-One and Small Group Tutoring

All students in both the individual tutoring and small group tutoring conditions in the MyST-MP&D study were administered a written questionnaire. Students in both groups received and responded to the same set of questions as those used in the MyST-SDS study, displayed above.

In addition, students in the small group condition each responded to questions that were designed to gain insights about students' experiences about working with other students in small groups. Results of the questionnaire indicate that students had quite similar impressions in the two conditions. Students in small groups indicated that they benefitted from group discussions, and interesting, indicated that they often disagreed with the answer that was provided by the designated speaker after the group discussion.

Figure 13 One-On-One vs Group Discussions

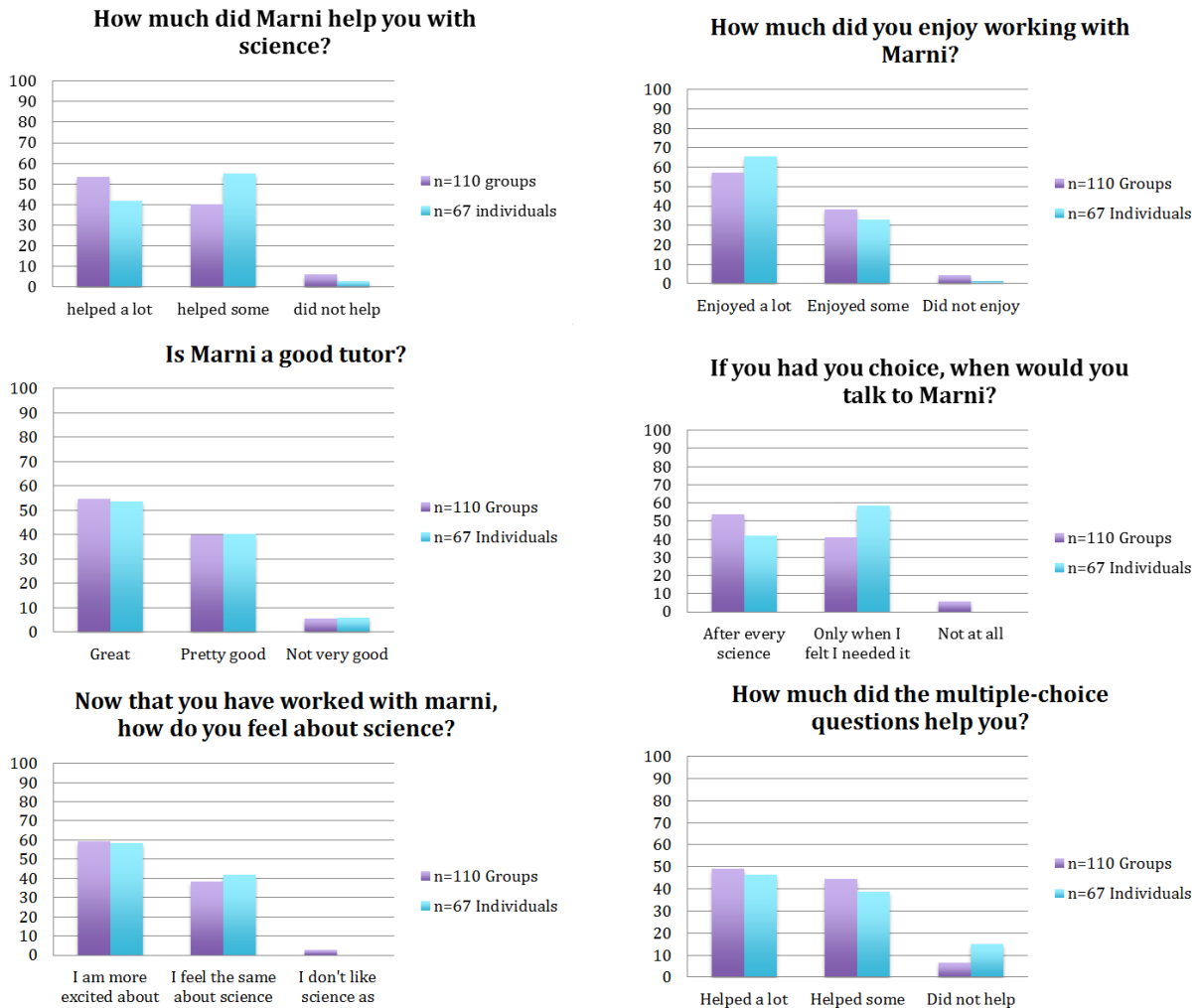


Figure 14 – Students in Small Groups Survey Results

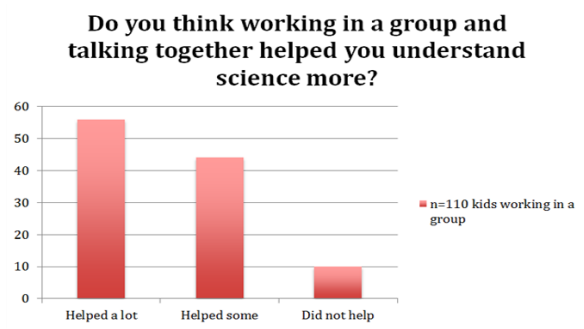
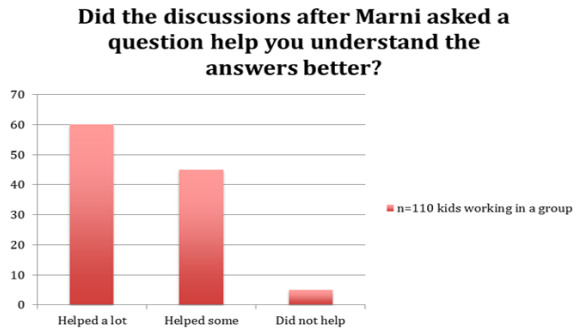
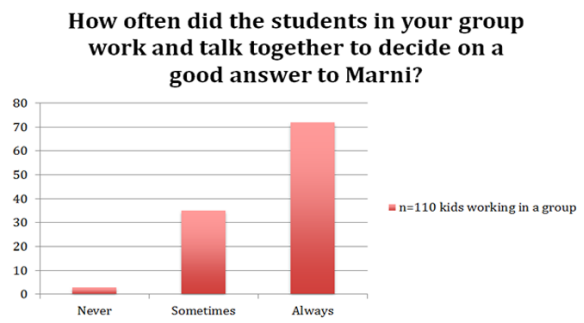
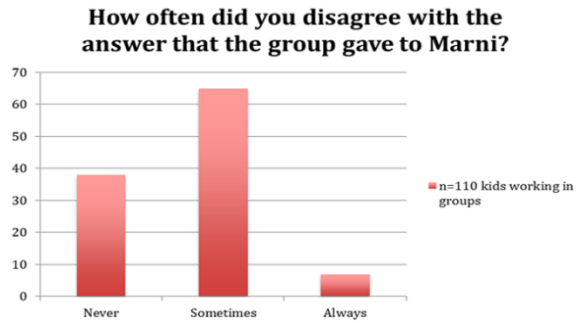
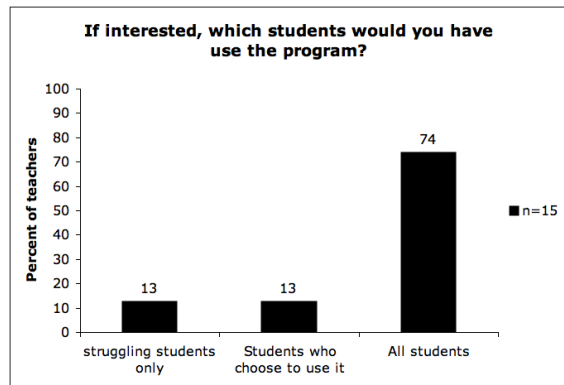
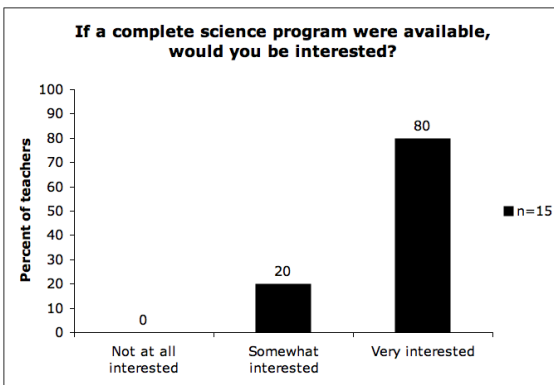
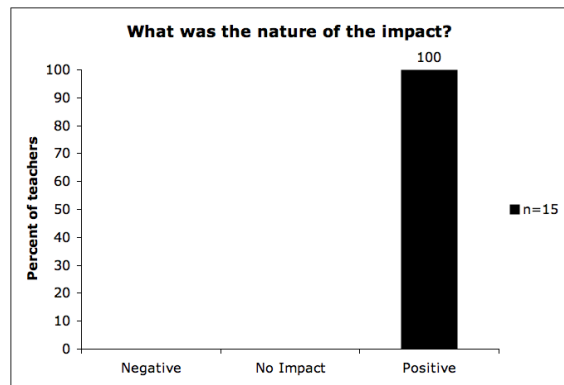
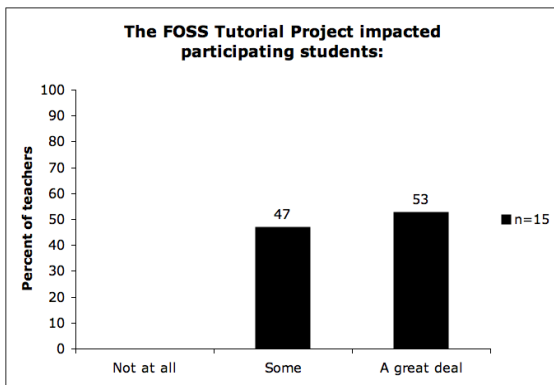
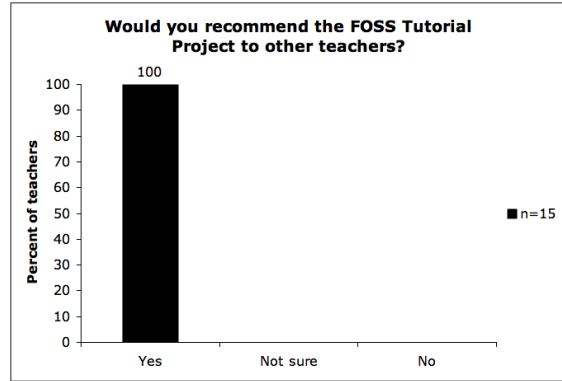
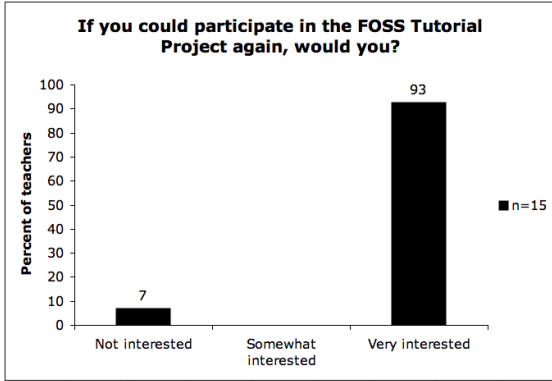


Figure 15: Combined Teacher Survey Results; MyST-SLS and MyST-MP&D





REFERENCES

- Baker, Good, Knutson, & Watson. (2006). *Indicadores dinámicos del éxito en la lectura* (7th ed.). Eugene, OR: Dynamic Measurement Group.
- Baker, Good, Mross, McQuilkin, Watson, Chaparro, & Sanford. (2006). Fluidez en la lectura oral idel *In indicadores dinámicos del éxito en la lectura*. Eugene, OR: Dynamic Measurement Group.
- Baker, Good, Peyton, & Watson. (2004). *Alternate form reliability of idel fluidez en las palabras sin sentido (raw data)*. Eugene, OR: University of Oregon.
- Baker, Park, & Baker. (2013). Effect of initial status and growth in pseudoword reading on spanish reading comprehension at the end of first grade. *Psicothema*.
- Baker, Park, Baker, & Basaraba. (2012). Effects of a paired bilingual reading program and an english-only program on the reading performance of english learners in grades 1–3 *Journal of School Psychology, 50*(6), 737–758.
- Baker, Smolkowski, Katz, Fien, Seeley, Kame'enui, & Beck. (2008). Reading fluency as a predictor of reading proficiency in low-performing high poverty schools. *School Psychology Review, 37*, 18-37.
- Baker, Smolkowski, Mielke, Linan-Thompson, Kosti, & Miciak. (inPreparation). Examining the effectiveness of systematic and explicit routines on spanish reading outcomes for first grade spanish-speaking english learners.
- Bandura, A. (1977). *Social learning theory*. New York, NY: General Learning Press.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive*. Englewood Cliffs, NJ: Prentice Hall.
- Barker, L. (2003). Computer-assisted vocabulary acquisition: The cslu vocabulary tutor in oral-deaf education. *Journal of Deaf Studies and Deaf Education, 8*(2), 187 - 198.
- Bazerman, C. (1998). *Shaping written knowledge*. Madison, WA: University of Wisconsin Press.
- Beck, & McKeown. (2006). *Improving comprehension with questioning the author: A fresh and expanded view of a powerful approach*: Scholastic.
- Beck, McKeown, Worthy, Sandora, & Kucan. (1996). Questioning the author: A year-long classroom implementation to engage students with text. *The Elementary School Journal, 96*(4), 387-416.
- Beck, I., McKeown, M., Sandora, C., Kucan, L., & Worthy, J. (1996). Questioning the author: A yearlong classroom implementation to engage students with text. *The Elementary School Journal, 96*(4), 385-414.
- Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In Gardner (Ed.), *Assessment and learning* (pp. 81-100). London: Sage.
- Bloom. (1984a). The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. *Educational Researcher, 13*, 4 - 16.
- Bloom. (1984b). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16.
- Brem, S., Bach, S., Kucian, K., Guttorm, T., Martin, E., Lyytinen, H., . . . Richardson, U. (2010). Brain sensitivity to print emerges when children learn letter-speech sound correspondences. *National Academy of Sciences (PNAS), 107*(17), 7939-7944.
- Bruner. (1985). Vygotsky: A historical and conceptual perspective. In Wertsch (Ed.), *Culture, communication, and cognition: Vygotskian perspectives* (pp. 21-34). Cambridge, England: Cambridge University Press.

- Cazden. (1979). Peekaboo as an instructional model: Discourse development at home and at school. *Stanford Papers and Reports in Child Language Development*, 17(1-19).
- Chaiklin. (2003). The zone of proximal development in vygotsky's analysis of learning and instruction. In Kozulin, Gindis, Ageyev & Miller (Eds.), *Vygotsky's educational theory and practice in cultural context*. Cambridge: Cambridge University Press.
- Chang-Wells, & Wells. (1993). Dynamics of discourse: Literacy and the construction of knowledge. In Forman, Minick & Stone (Eds.), *Contexts for learning: Sociocultural dynamics in children's development* (pp. 58-90). New York, NY: Oxford University Press.
- Chi, Bassok, Lewis, Reimann, Glaser, & Alexander. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2), 145-182.
- Chi, DeLeeuw, Chiu, & LaVancher. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 439-477.
- Chi, Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science*, 25, 471-533.
- Cobb, & Yackel. (1996). Constructivist, emergent, and sociocultural perspectives in the context of developmental research. *Educational Psychologist*, 31(3-4).
- 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, & Engeström. (1993). A cultural-historical approach to distributed cognition. In Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations* (pp. 1-46). New York, NY: Cambridge University Press.
- Cole, Halpern, Ramig, Vuuren, v., Ngampatipatpong, & Yan. (2007). A virtual speech therapist for individuals with parkinson disease. *Educational Technology*, 47(1), 51-55.
- Cole, Vuuren, V., Pellom, Hacıoglu, Ma, Movellan, . . . Yan. (2003). Perceptive animated interfaces: First steps toward a new paradigm for human-computer interaction. *Proceedings of the IEEE*, 91(9), 1391-1405.
- Cole, Wise, & Vuuren, V. (2007). How marni teachers children to read. *Educational Technology*, 24(1), 14-18.
- Cole, M. (1996). *Cultural psychology*. Cambridge, MA: Harvard University Press.
- Coyne, Kame'enui, & Carnine. (2011). *Effective teaching strategies that accommodate diverse learners*. Upper Saddle River, NJ: Pearson.
- Davis. (2004). Explorations of scaffolding in complex classroom systems. *Journal of the Ledaarning Sciences*, 13(3), 265-272.
- Davis, E., & Miyake, N. (2004). Explorations of scaffolding in complex classroom systems. *Journal of the Learning Sciences*, 13(3), 265-272.
- deCara, B., & Goswami, U. (2002). Statistical analysis of similarity relations among spoken words: Evidence for the special status of rimes in english. *Behavioural Research Methods and Instrumentation*, 34(3), 416-423.
- Driscoll, D., Craig, S., Gholson, B., Ventura, M., Hu, X., & Graesser, A. (2003). Vicarious learning: Effects of overhearing dialog and monologue-like discourse in a virtual tutoring session. *Journal of Educational Computing Research*, 29(4), 431-450.
- Fernald, Marchman, & Weisleder. (2013). Ses differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, 16(2), 234-248.

- Fien, Baker, Smolkowski, Smith, Kame'enui, & Beck. (2008). Using nonsense word fluency to predict reading proficiency in kindergarten through second grade for english learners and native english speakers. *School Psychology Review*, 37(3), 391-408.
- FOSS. (2007). from <http://www.fossweb.com>
- Foucault, M. (1969). *The archeology of knowledge*. New York, NY: Random House.
- Fuchs, L.S., Fuchs, D., Hosp, M.K., & Jenkins, J.R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5, 239-256.
- Gee, J.P. (1990). *Social linguistics and literacies*. London: Falmer Press.
- Geertz, C. (1983). *Local knowledge*. New York, NY: Basic Books.
- Gholson, B., Witherspoon, A., Morgan, B., Brittingham, J.K., Coles, R., Graesser, A.C., . . . Craig, S.D. (2009). Exploring the deep-level reasoning questions effect during vicarious learning among eighth to eleventh graders in the domains of computer literacy and newtonian physics. *Instructional Science*, 37(5), 487-493.
- Good, Baker, & Peyton. (2009). Making sense of nonsense word fluency: Determining adequate progress in early first-grade reading. *Reading & Writing Quarterly*, 25, 33-56.
- Good, & Kaminski. (2002). *Dynamic indicators of basic early literacy skills (6th ed.)*. Eugene, OR: Inisitute for the Development of Educational Achievement.
- Good, Simmons, & Kame'enui. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, 5, 257-288.
- Good, Wallin, Simmons, Kame'enui, & Kaminski. (2002). Systemwide percentile ranks for dibels benchmark assessment. Eugene, OR: University of Oregon.
- Gough, Hoover, & Patterson. (1996). Some observations on a simple view of reading. In C. Cornoldi, Oakhill (Ed.), *Reading comprehension difficulties: Processes and intervention* (pp. 1-13). Mahway, New Jersey: Lawrence Erlbaum Associates.
- Gough, & Tunmer. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7, 6-10.
- Graesser, A.C., & Person, N.K. (1994). Question asking during tutoring. *American Educational Research Journal*, 31(1), 104-107.
- Halliday, M. (1978). *Language as social semiotic*. London: Edward Arnold.
- Haraway, D. (1989). *Primate visions*. New York, NY: Routeledge.
- Haraway, D. (1991). *Simians, cyborgs, and women*. New York, NY: Routeledge.
- Haraway, D. (1999). *Modest witness @ second millennium*. New York, NY: Routeledge.
- Harcourt, H.M. (2010). from <http://www.hmhco.com/shop/education-curriculum/reading/core-reading-programs/journeys>
- Harland. (2003). Vygotsky's zone of proximal development and problem-based learning: Linking a theoretical concept with practice through action research. *Teaching in Higher Education*, 8(2), 263-272.
- Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday lives of young american children*. Baltimore, MD: Brookes.
- Hausmann, & VanLehn. (2007). Explaining self-explaining: A contrast between content and generation. In Luckin, Koedinger & Greer (Eds.), *Artificial intelligence in education* (pp. 417-424). Amsterdam, Netherlands: IOS Press.

- Hausmann, & VanLehn. (2007b). *Self-explaining in the classroom: Learning curve evidence*. Paper presented at the 29th Annual Conference of the Cognitive Science Society, Mahwah, NJ.
- Higgins, Hartley, & Skelton. (2002). The conscientious consumer: Reconsidering the role of assessment feedback in student learning. *Studies in Higher Education*, 27(1), 53-64.
- Holbrook, & Kolodner. (2000). *Scaffolding the development of an inquiry-based (science) classroom*. Paper presented at the Fourth International Conference of the Learning Sciences, Mahwah, NJ.
- Hoover, & Gough. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, 2, 127-160.
- Hutchins, E. (1980). *Culture and inference*. Cambridge, MA: Harvard University Press.
- John-Steiner, & Mahn. (1996). Sociocultural approaches to learning and development: A vygotskian framework. *Educational Psychologist*, 31(3/4), 191-206.
- John-Steiner, Panofsky, & Smith. (1994). *Sociocultural approaches to language and literacy: An interactionist perspective*. New York, NY: Cambridge University Press.
- King. (1991). Effects of training in strategic questioning on children's problem-solving performance. *Journal of Educational Psychology*, 83, 307-317.
- King. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*, 31(2), 338.
- King, Staffieri, & Adalgais. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology*, 90(1), 134-152.
- Kirschner, Sweller, & Clark. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75-86.
- Kohler, E., C., K., Umiltà-Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science*, 297, 846-848.
- Kujala, T., Lovio, R., Halttunen, A., Lyytinen, H., & Näätänen, R. (2012). Reading skill and neural processing accuracy improvement after a 3-hour intervention. In preschoolers with difficulties in reading-related skills. *Brain Research*, 1448, 42-55.
- Kyle, F., Kujala, J., Richardson, U., Lyytinen, H., & Goswami, U. (2013). Assessing the effectiveness of two theoretically motivated computer-assisted reading interventions in the united kingdom: Gg rime and gg phoneme. *Reading Research Quarterly*, 48(1), 61-76.
- LaBerge, D., & Samuels, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293-323.
- Latour, B. (1987). *Science in action*. Cambridge, MA: Harvard University Press.
- Lave, J. (1988). *Cognition in practice*. Cambridge, UK: Cambridge University Press.
- Lee, L., & Rose, R.C. (1998). A frequency warping approach to speaker normalization. *IEEE Trans. Speech Audio Process.*, 6(1), 49-60.
- Leggetter, C.J., & Woodland, P.C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9, 171-185.
- Lemke. (2001). Articulating communities: Sociocultural perspectives on science education. *Journal of Research in Science Teaching*, 38(3), 296-316.

- Lemke. (2006). Towards critical multimedia literacy: Technology, research, and politics. In McKenna, Reinking, Labbo & Kieffer (Eds.), *International handbook of literacy & technology*, v2.0. Mahwah, NJ: Erlbaum.
- Lemke. (2012). Multimedia and discourse analysis. In Gee & Handford (Eds.), *Routledge handbook of discourse analysis*.
- Lemke, J. (1990). *Talking science: Language, learning, and values*. Norwood, NJ: Ablex.
- Lemke, J. (1995). *Textual politics*. London: Taylor and Francis.
- Lemke, J. (1998a). Multimedia literacy demands of the scientific curriculum. *Linguistics and Education*, 10(3), 247-271.
- Lemke, J. (1998b). Multiplying meaning: Visual and verbal semiotics in scientific text. In J. R. Martin & R. Veel (Eds.), *Reading science* (pp. 87-113). London: Routledge.
- Lynch, M., & Woolgar, S. (1990). *Representation in scientific practice*. Cambridge, MA: MIT Press.
- Lyons. (1984). Defining a child's zone of proximal development: Evaluation process for treatment planning. *American Journal of Occupational Therapy*, 38(446-451).
- Madden, N.A., & Slavin, R.E. (1989). Effective pullout programs for students at risk. In R. E. Slavin, N. L. Karweit & N. A. Madden (Eds.), *Effective programs for students at risk*. Boston, MA: Allyn and Bacon.
- Martin, J. (1992). *English text*. Philadelphia, PA: John Benjamins.
- Mayer. (2001). *Multimedia learning*. Cambridge, UK: Cambridge University Press.
- Mayer. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and instruction*, 13(2), 125-139.
- Mayer. (2005). The cambridge handbook of multimedia learning (pp. 169-182). New York, NY: Cambridge University Press.
- McKeown, M., & Beck, I. (1999). Getting the discussion started. *Educational Leadership*, 57(3), 25-28.
- McKeown, M., Beck, I., Hamilton, R., & Kucan, L. (1999). *"Questioning the author" accessibles: Easy access resources for classroom challenges*. Bothell, WA: The Wright Group.
- McNamara, Levinstein, & Boonthum. (2004). Istart: Interactive strategy training for active reading and thinking. *Behavioral Research Methods, Instruments, and Computers*(36), 222-233.
- Mishler, E. (1984). *The discourse of medicine*. Norwood, NJ: Ablex.
- Morgan. (2013). *Science achievement gaps in the us: A longitudinal investigation*. Paper presented at the Children's Learning Research Collaborative (CLRC), Ohio State University.
- Mostow, J., & Aist, G. (1999). Giving help and praise in a reading tutor with imperfect listening because automated speech recognition means never being able to say you're certain. *CALICO Journal*, 16(3), 407-424.
- Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of project listen. In K. Forbus & P. Feltoich (Eds.), *Smart machines in education* (pp. 169-234). Menlo Park, CA: MIT/AAAI Press.
- Murphy, P., Wilkinson, I., Soter, A., Hennessey, M., & Alexander, J. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology*, 101(3), 740-764.

- NAEP, N.A.o.E.P. (2005). National and state reports in science *The Nations Report Card: National Assessment of Educational Progress*.
- National Research Council. (1999). How people learn: Brain, mind, experience, and school. In J. D. Bransford, A. L. Brown & R. R. Cocking (Eds.), *Committee on Developments in the Science of Learning*. Washington, DC: The National Academies Press.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press.
- National Research Council. (2007). Taking science to school: Learning and teaching science in grades k-8. In R. A. Duschl, H. A. Schweingruber & A. W. Shouse (Eds.), *Committee on Science Learning Kindergarten through Eighth Grade*. Washington D.C.: The National Academies Press.
- National Research Council. (2011a). *A framework for k-12 science education: Practices, crosscutting concepts, and core ideas*: The National Academies Press.
- National Research Council. (2011b). Successful k-12 stem education: Identifying effective approaches in science, technology, engineering, and mathematics *Committee on Highly Successful Science Programs for K-12 Science Education. Board on Science Education and Board on Testing and Assessment*. Washington, DC: Division of Behavioral and Social Sciences and Education.
- National Research Council. (2013). Developing assessments for the next generation science standards. In C. o. D. A. o. S. P. i. K.-B. o. T. a. A. B. o. S. Education (Ed.). Washington, DC: Behavioral and Social Sciences and Education.
- NRC. (2011). A framework for k-12 science education: Practices, crosscutting concepts, and core ideas.
- Nystrand, & Gamoran. (1991). Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*, 25(3), 261-290.
- Nystrand, Gamoran, Kachur, & Prendergast. (1997). *Opening dialogue: Understanding the dynamics of language and learning in the english classroom*. New York, NY: Teachers College Press.
- Nystrand, M., Gamoran, A. (1991). Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*, 25(3), 261-290.
- Obukhova, & Korepanova. (2009). The zone of proximal development: A spatiotemporal model. *Journal of Russian & East European Psychology*, 47(6), 25-47.
- Ojanen, E., Kujala, J., Richardson, U., & Lyytinen, H. (2013). Technology enhanced literacy learning in zambia. *Insights on Learning Disabilities*, 10(2), 103.
- Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, 328, 463-466.
- Palincsar, & Brown. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1(2), 117-175.
- Palincsar, & Brown. (1986). Interactive teaching to promote independent learning from text. *The Reading Teacher*, 39, 771-777.
- Palincsar, & Brown. (1988). Teaching and practicing thinking skills to promote comprehension in the context of group problem solving. *Remedial and Special Education (RASE)*, 9(1), 53-59.
- Pea. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences*, 13(3), 423-451.

- Perfetti, C. (1985). *Reading ability*. Oxford, England: Oxford University Press.
- Pine, & Messer. (2000). The effect of explaining another's actions on children's implicit theories of balance. *Cognition and Instruction*, 18(1), 35-52.
- Puntambekar, & Hübscher. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist*, 40(1), 1-12.
- Puntambekar, & Kolodner. (2005). Distributed scaffolding: Helping students learn science by design. *Journal of Research in Science Teaching & Learning in Medicine*, 42.
- Reid. (1998). Scaffolding: A broader view. *Journal of Learning Disabilities*, 31, 386-396.
- Reynolds, R.E. (2000). Attentional resource emancipation: Toward understanding the interaction of word identification and comprehension processes in reading. *Scientific Studies of Reading*, 4, 169-195.
- Rizzolatti, G., & Craighero, L. (2007). Language and mirror neurons. In Gaskell (Ed.), *The oxford handbook of psycholinguistics* (pp. 771-785). Oxford: Oxford University Press.
- Roehler, & Cantlon. (1997). Scaffolding: A powerful tool in social constructivist classrooms. In Hogan & Pressley (Eds.), *Scaffolding student learning: Instructional approaches and issues* (pp. 6-42). Cambridge, MA: Brookline.
- Rogoff. (1994). Developing understanding of the idea of communities of learners. *Mind, Culture, and Activity*, 1, 209-229.
- Rogoff. (1999). Thinking and learning in a social context. In Lave (Ed.), *Everyday cognition: Development in social context* (pp. 1-8). Cambridge, MA: Harvard University Press.
- Rogoff, B. (1990). *Apprenticeship in thinking*. New York, NY: Oxford University Press.
- Sampson, V., & Grooms, J. (2010). Promoting and supporting scientific argumentation in the classroom: The generate an argument instructional model. *The Science Teacher*, 77(5), 33-37.
- Samuels, S. (1997). The importance of automaticity for developing expertise in reading. *Reading and Writing Quarterly* 13, 107-122.
- Schegloff, E. (1991). Reflections on talk and social structure. In D. Boden & D. Zimmerman (Eds.), *Talk and social structure* (pp. 44-70). Berkeley, CA: University of California Press.
- Shapin, S., & Schaffer, S. (1985). *Leviathan and the air-pump*. Princeton, NJ: Princeton University Press.
- Smith. (1941). *Measurement of the size of general english vocabulary through the elementary grades and high school*.
- Soter, A., Wilkinson, I., Murphy, P., Rudge, L., Reninger, K., & Edwards, M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research*, 47, 372-391.
- Spindler, G. (1987). *Education and cultural process (2nd edition)*. Prospect Heights, IL: Waveland Press.
- Stanovich, K.E. (2000). The interactive-compensatory model of reading: A confluence of developmental, experimental, and educational psychology. In K. E. Stanovich (Ed.), *Progress in understanding reading: Scientific foundations and new frontiers* (pp. 44-54). New York, NY: Guilford Press.
- Sullins, J., Craig, S.D., & Graesser, A.C. (2010). The influence of modality of deep reasoning questions. *International Journal of Learning Technology*, 5, 378-387.
- The Future of Children. (2005). *School Readiness: Closing Racial and Ethnic Gaps* 15(1).

- Topping, K., & Whiteley, M. (1990). Participant evaluation of parent-tutored and peer-tutored projects in reading. *Educational research*, 32(1), 14-32.
- VanLehn. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. *Educational Psychologist*, 46(4), 197-221.
- VanLehn, Graesser, Jackson, Jordan, Olney, & Rose. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science* 31(1), 3-62.
- VanLehn, K., & Graesser, A. (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*.
- Vaughn, Cirino, Linan-Thompson, Mathes, Carlson, Hagan, . . . Francis. (2006). Effectiveness of a spanish intervention and an english intervention for english-language learners at risk for reading problems. *American Education Research Journal*, 43(3), 449-487.
- Vygotsky. (1962). *Thought and language*. Cambridge, MA: MIT Press.
- Vygotsky. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Vygotsky. (1981). The instrumental method in psychology. In J. V. Wertsch (Ed.), *The concept of activity in soviet psychology* (pp. pg. 134-144). Armonk, NY: M.E. Sharpe.
- Vygotsky. (1986). *Thought and language*. Cambridge, MA.
- Vygotsky. (1987). *Thinking and speech*. New York, NY: Plenum.
- Ward, Cole, Bolanos, Buchenroth-Martin, Svirsky, vanVuuren, . . . Becker. (2011). My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Trans. Speech Lang. Process.*, 7(4), 18.
- Ward, Cole, Bolanos, Buchenroth-Martin, Svirsky, & Weston. (2013). My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology*, 105(4), 1115-1125.
- Ward, W., Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S. V. (2011). My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Trans. Speech Lang. Process.*, 7(4).
- Watson, K., & Froyd, J. (2007). Diversifying the u.S. Engineering workforce: A new model. *Journal of Engineering Education*, 96(1), 19-32.
- Wells. (1994). *Changing schools from within: Creating communities of inquiry*. Portsmouth, NH: Toronto: OISE Press.
- Wells. (1999). *Dialogic inquiry: Towards a sociocultural practice and theory of education*. New York, NY.
- Wells. (2000). Dialogic inquiry in education: Building on the legacy of vygotsky. In Lee & Smagorinsky (Eds.), *Vygotskian perspectives on literacy research* (pp. 51-85). New York, NY: Cambridge University Press.
- Wertsch. (1984). The zone of proximal development: Some conceptual issues. In Rogoff & Wertsch (Eds.), *New directions for child development: No. 23. Children's learning in the "zone of proximal development*. San Francisco: Jossey-Bass.
- Wertsch, J.V. (1985). *Vygotsky and the social formation of mind*. Cambridge, MA: Harvard University Press.
- Wertsch, J.V. (1991). *Voices of the mind: A sociocultural approach to mediated action*. Cambridge, MA: Harvard University Press.

Wilson, & Weinstein. (1996). The transference and the zone of proximal development. *Journal of the American Psychoanalytic Association*, 44, 167-200.

APPENDIX A

MyST's Theoretical and Empirical Foundations

In this appendix we review the theoretical foundations and scientific rationale for the design decisions and dialog strategies in the MyST systems.

We note that MyST was influenced by a series of NRC reports (National Research Council, 1999, 2001, 2007, 2011a, 2011b, 2013). Foremost among these was "Taking Science to School: Learning and Teaching Science in Grades K-8" (National Research Council, 2007). This report emphasizes the critical importance of scientific discourse in K-12 science education, and highlights crucial principles of scientific proficiency: "Students who are proficient in science: 1. know, use, and interpret scientific explanations of the natural world; 2. generate and evaluate scientific evidence and explanations; 3. understand the nature and development of scientific knowledge; and 4. participate productively in scientific practices and discourse." (pg. 2).

The report also emphasized that *scientific inquiry and discourse is a learned skill*, so students need to be involved in activities in which they learn appropriate norms and language for productive participation in scientific discourse and argumentation. MyST-SDS was designed to help students and to achieve proficiency in scientific discourse, reasoning and argumentation. These skills must be acquired for students to achieve proficiency in science learning in U.S. classrooms, consistent with the Next Generation Science Standards (NGSS, 2013) and the recommendations of the NRC (2011) report, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. The new MyST dialogs developed with support from the IES Goal 3 grant, incorporate science content that is aligned to the NGSS.

1. Sociocultural Perspectives on Learning

MyST and CASUM are based on sociocultural views of learning. Lemke (REF) provides a concise summary of the development of the sociocultural movement:

Lemke (2001) presents an excellent historical perspective on sociocultural influences on science education:

"The view that science represents a uniquely valid approach to knowledge, disconnected from social institutions, their politics, and wider cultural beliefs and values was strongly challenged by research in the history of science (Shapin & Schaffer, 1985), the sociology of science (Latour, 1987; Lynch & Woolgar, 1990), and ethnoscience studies in cultural anthropology (Hutchins, 1980), and contemporary science studies (Haraway, 1989, 1991, 1999). Historians, sociologists, and cultural anthropologists came increasingly to see that science had to be understood as a very human activity whose focus of interest and theoretical dispositions in any historical period were, and are, very much a part of, and not apart from the dominant cultural and political issues of the day." (p. 300).

"... the view of science education (and education in general) as a second socialization or specialist enculturation into a sub-community was developed out of anthropological theory (Lave, 1988; Spindler, 1987) and neo-Vygotskian perspectives in developmental psychology (M. Cole, 1996; B. Rogoff, 1990; J. V. Wertsch, 1991) in opposition to asocial views of autonomous cognitive development." (pg. 300)

“Finally, along with all the social sciences in this period (Foucault, 1969; Geertz, 1983), both science education and the new science studies (in history and sociology) took the ‘linguistic turn’ and began to examine how people learned to talk and write the languages of science and meaningfully and cooperatively engage in its wide range of subculturally specific activities (e.g. observing, experimenting, publishing) and signifying practices (data tabulation, graphing, etc.). In place of a Chomskyan view of language as an automatic, gene-guided machine for correct syntax, people who were studying the functions of language in social interaction (Bazerman, 1998; Halliday, 1978; J. Lemke, 1990; Martin, 1992; Mishler, 1984; Schegloff, 1991) began to see language as a culturally transmitted resource for making meaning socially (Gee, 1990; J. Lemke, 1995) that was also useful for talking oneself through science problems. Language, however, was just one such tool; science and science learning are in fact best characterized by their rich synthesis of linguistic, mathematical, and visual representations (J. Lemke, 1998a, 1998b; Lynch & Woolgar, 1990) In the sociocultural view, what matters to learning and doing science is primarily the socially learned cultural traditions of what kinds of discourses and representations are useful and how to use them, far more than whatever brain mechanisms may be active while we are doing so.” (Page 301)

Vygotsky and Social Constructivism

Lev Vygotsky’s writings have profoundly influenced educational research, classroom instructional treatments, and intelligent tutoring system in the US and worldwide. Social constructivism holds that all learning is culturally embedded and socially mediated. Knowledge is acquired in social contexts and is mediated by language. (Vygotsky, 1962, 1978, 1981, 1986, 1987; J. V. Wertsch, 1985). In Vygotsky’s view, language and thought were inseparable and synergistic:

“The relation of thought to word is not a thing but a process, a continual movement backward and forth from thought to word and from word to thought. In that process, the relation of thought to word undergoes changes that themselves may be regarded as developmental in the functional sense. Thought is not merely expressed in words; it comes into existence through them. Every thought tends to connect something with something else, to establish a relation between things. Every thought moves, grows and develops, fulfills a function, solves a problem.” (1986, p.218)”

Vygotsky’s views on learning and language greatly influenced our conceptualization and implementation of MyST dialog strategies, as well as the design of CASUM dialogs. These effects are both direct, i.e., based on Vygotsky’s writings, and indirect, as his work influenced many prominent theorists and researchers who have applied his ideas to classroom programs and intelligent tutoring systems.

Science is special: It is interesting to note that Vygotsky (1987) believed that the acquisition of scientific vocabulary and knowledge differed in fundamental ways from the “spontaneous” or “everyday” acquisition of word meanings and knowledge. Whereas Vygotsky believed that the acquisition of word meanings during everyday conversations was based on the social contexts in which they occurred, he wrote that scientific terms were learned initially through definitions provided by teachers, and that learning science required learning the precise meanings of words and their relationships to each other within specific scientific systems. Thus, while it is still the case that students’ prior experiences will strongly influence what they hear and understand, and

how others interpret what they say during classroom science instruction, it is also the case that *all students must learn the language of scientific discourse and argumentation, which has its own rules and conventions*. We this idea in focus when developing MyST dialogs. The fact that all students must learn specific norms and vocabulary to engage in scientific discourse creates a more level playing field for all students. It makes tutoring within MyST tractable and achievable, since the virtual tutor can model and reinforce appropriate use of science vocabulary and discourse for all students.

Vygotsky defined the **Zone of Proximal Development**, or ZPD, as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers" (Vygotsky, 1978) (pg. 86). Vygotsky viewed the ZPD as the zone in which learning can be optimized, since learners can be stimulated to integrate new information with prior knowledge to construct new knowledge. One implication of keeping students in the ZPD is that they must master foundational knowledge, e.g., vocabulary and concepts, so they can build on this knowledge, with help from a teacher or more competent peer, to construct new knowledge.

According to John-Steiner and Mahn (1996), "Sociocultural theorists, expanding the concept of the zone of proximal development, increasingly conceptualize learning as distributed (Cole & Engeström, 1993), interactive (Chang-Wells & Wells, 1993), contextual (John-Steiner, Panofsky, & Smith, 1994), and the result of the learners' participation in a community of practice (Chang-Wells & Wells, 1993; Cole & Engeström, 1993; John-Steiner & Mahn, 1996; John-Steiner et al., 1994; Rogoff, 1994). Numerous authors have discussed Vygotsky's ideas about the ZPD and ways to measure the ZPD in learning and education (Chaiklin, 2003; Harland, 2003; Lyons, 1984; Obukhova & Korepanova, 2009; Wertsch, 1984; Wilson & Weinstein, 1996).

Scaffolding is the process by which teachers or tutors stimulate and challenge students to construct new knowledge in the ZPD by providing them with new information—including questions, hints, drawings, or gestures—that they can use to construct new knowledge. The term "scaffolding" which was described by Vygotsky but was not appear in his writings, has become commonplace in describing the process of providing new information that stimulates and motivates children to learn within the ZPD. Vygotsky suggested that teachers use cooperative learning exercises in which more knowledgeable students can work with their less knowledgeable peers to facilitate learning within the ZPD. Several publications have discussed the importance of and means for using scaffolds effectively in classroom instructional treatments (Davis, 2004; E. Davis & Miyake, 2004; Holbrook & Kolodner, 2000; Pea, 2004; Puntambekar & Hübscher, 2005; Puntambekar & Kolodner, 2005; Reid, 1998; Roehler & Cantlon, 1997).

Reciprocal tutoring is an example of an instructional approach inspired by Vygotsky's work (King, Staffieri, & Adelgais, 1998; Palincsar & Brown, 1984, 1986, 1988). In this approach a teacher works with groups of students to discuss a text passage. The teacher models how to explain ideas presented in the text, and each student learns to play the role of group leader who explains the text passage to other students. Reciprocal learning has been shown to improve students' engagement, motivation and text comprehension (Palincsar & Brown, 1984). VanLehn (2011); VanLehn et al. (2007); K. VanLehn and Graesser (2002) identified scaffolding as one of two strategies that are most effective in accounting for the consistent and substantial

learning gains obtained in a large number of studies which human tutoring and intelligent tutoring systems to classroom instruction.

In sum, Vygotsky's writings led to new perspectives on the relationship between culture, language, thinking and learning that has had a profound influence on the writings and research of prominent philosophers and researchers, and stimulated research that has guided approaches to human tutoring, the design of intelligent tutoring systems and the design of classroom instructional approaches (Bruner, 1985; Cazden, 1979; Cobb & Yackel, 1996; John-Steiner & Mahn, 1996; Rogoff, 1999; B. Rogoff, 1990; Wells, 1994, 1999, 2000). Below, we discuss research related to the specific dialogs strategies used in MyST.

Home Environments, School Readiness and School Achievement

An implication of sociocultural views is that, if knowledge is acquired in social contexts using language, then the quality of children's early social and linguistic experiences should have a profound effect on their knowledge acquisition and language proficiency, and the ways in which they engage in social interactions. Children who grow up in homes where they are engaged in a wide range of child-centered conversations, where language is varied and used creatively, where parents and children interact with resources such as books or educational software, they will arrive at school with more knowledge, language proficiency, social awareness and self-efficacy than students who do not have these home experiences.

Over 70 years of research supports the following conclusions: a) children who live in lower-income homes with less educated parents are likely to enter school with poorer language skills than their more privileged peers, b) children's language skills when they enter school, measured by their vocabulary knowledge, is a strong predictor of future academic success, and c) it is extremely difficult to close the achievement gap between children with poor language skills and their higher performing peers. The evidence in support of each of these conclusions is compelling.

Smith (1941) administered an English vocabulary test to students in first grade through high school. Results showed that "high knowledge third graders had vocabularies about equal to lowest-performing 12th graders" and that "high-school seniors near the top of their class knew about four times as many words as their lower-performing classmates".

Hart and Risley (1995) recorded the language of professional, working class and welfare families in their homes in Kansas during a period of 2 and a half years. Children from welfare families heard, on average, 616 words per hour, whereas children from professional families heard 2153 words per hour. Longitudinal studies of these children revealed a high correlation between vocabulary knowledge at age three and language proficiency and academic success at ages nine and ten.

Morgan (2013) conducted an analysis of the Early Childhood Longitudinal Study, a sample of U.S. Kindergarten Class of 1998-99, which assessed a representative sample of U.S. Students entering school in 1998 that were tested for their science, math and reading achievement in kindergarten, first, third and eighth grades (1998, 2000, 2002, 2007.) He compared student achievement as a function of students' race/ethnicity, parents' marital status, mother's educational level and family income. These factors accounted for between 70% to 80% of the variance in children's science achievement through eighth grade. Morgan concluded: "The study's modeling fully explained the Hispanic-, American Indian-, and Asian-White science

achievement gaps by 8th grade, and mostly explained the Black-White science achievement gap.” Moreover, “Early, constrained opportunities and propensities to learn science, reading, and mathematics in the preschool period, lower learning-related behavioral functioning, and social class characteristics largely explain science achievement gaps between racial/ethnic minorities in the U.S.”

Effects of home environment on children’s language processing have been demonstrated as early as 18 months of age. Fernald, Marchman, and Weisleder (2013) found that toddlers from disadvantaged families are already several months behind more advantaged children in language proficiency (Fernald et al., 2013). Toddlers were presented with a pair of objects, and asked to look at one of them. Children from homes with low SES poorer were 200 milliseconds slower than children from middle class homes in their response times.

Results of the NAEP (2005) highlight the differences in academic achievement of children in U.S. elementary and middle schools from different home environments based on SES, race and ethnicity. Students who are Black, Hispanic, or American Indian have lower science achievement than White students. For example, 50th percentile scores of Hispanics and American Indians fall below the 25th percentile scores of Whites in 4th and 8th grade, while the 50th percentile scores of Blacks approximate the 10th percentile scores of Whites.

For an informative, multidisciplinary treatment of the effects of home environment on school readiness and academic achievement, we recommend the collection of articles in the journal *Future of Children: School Readiness: Closing Racial and Ethnic Gaps* (The Future of Children, 2005).

2. Sociocultural Perspectives & Empirical Foundations of MyST Dialogs

This section identifies each of the dialog strategies or moves that were used by Marni, and provides empirical evidence that motivates their use.

Marni asks students authentic, deep reasoning questions. Marni's open-ended questions are designed both to model scientific discourse and scaffold learning, along with the media that accompanies the questions. These questions are designed to stimulate students to reason about and explain science. Marni never asks a question that had an obvious answer, such as "Which part of this circuit stores the electricity?" Instead, she might show a picture of a circuit and ask questions like: "So what's going on here?" "What's this all about?" As the dialog progresses, Marni's open-ended questions became more focused. "What else can you tell me about the direction of the flow of electricity?"

A significant body of research indicates that learning improves when teachers, tutors or students ask authentic, deep-reasoning questions (Graesser & Person, 1994; King, 1991; Murphy et al., 2009; Osborne, 2010; Sampson & Grooms, 2010; Soter et al., 2008). For example, when teachers read text passages to students, and then lead classroom conversations in which they ask authentic questions about the texts, students improve their comprehension of texts and their ability to engage in classroom discourse (Beck & McKeown, 2006; Beck, McKeown, Worthy, Sandora, & Kucan, 1996). Nystrand and Gamaron (1991) found that authentic dialogs, although rare in the classrooms studied, were most often initiated by authentic questions asked by students.

Marni models scientific discourse and appropriate use of scientific vocabulary. Marni typically initiates follow-on questions by first rephrasing parts of the students' previous answer. Thus, when talking about the flow of electricity in a circuit, if the student said "I see that it flows one way," Marni may respond, "I think I heard you say that the electricity flows through the circuit in one direction." A great deal of research has demonstrated that observing and modeling others' behaviors facilitates learning (Bandura, 1977, 1986). Recent research suggests that our brain's mirror neuron system plays a significant role in language learning; this system produces that mirror the behaviors of individuals we observe; the research indicates that we neural processes that help us recall and learn to produce language when we listen to and observe others speaking (Kohler, C., Umiltà-Fogassi, Gallese, & Rizzolatti, 2002; Rizzolatti & Craighero, 2007).

MyST continuously assesses students' understanding of the science being discussed. MyST dialogs are structured as a set of turns between Marni and the student; Marni asks the student a question, the student produces a spoken answer, and the spoken dialog system processes the answer to determine which concepts (represented as propositions within the dialog system) the student has expressed, and which remain to be expressed. The system then specifies the next question Marni will produce (which may be accompanied by new media); with the goal of helping students construct complete and accurate explanations. Continuously assessing students' level of science understanding of specific concepts enables the system to make judgments about whether students have mastered science concepts that serve as the foundation for new learning.

MyST helps students master prerequisite knowledge: MyST attempts to assure mastery of prior content by having students construct explanations that cover all of the points of each mini-dialog. However, if this does not occur after a specified number of dialog turns, MyST concludes the mini-dialog session. At this conclusion of each mini-dialog, Marni provides a

concise explanation of the key concepts of the learning goals of the mini-dialog. The spoken explanation, which incorporates media, is intended to help students construct an accurate multimodal (verbal and visual) understanding of the key concepts, consistent with the literature on multimedia learning reviewed below, so they can build on these concepts to learn new ones.

Acquisition of prerequisite knowledge is essential for subsequent learning of complex concepts. It is too often the case that teachers and even experienced tutors erroneously assume that students have mastered foundational knowledge that is a prerequisite for learning more advanced concepts. Bloom (1984b)'s seminal research on the benefits of classroom instruction verses tutoring revealed that one sigma gains could be obtained in classroom instruction by assuring that students master prior content before being introduced to new content that depends upon it.

Kirschner, Sweller, and Clark (2006) stress the critical importance of assuring that learners master prior content: “both the structures that constitute human cognitive architecture and evidence from empirical studies over the past half-century consistently indicate that minimally guided instruction is less effective and less efficient than instructional approaches that place a strong emphasis on guidance of the student learning process. The advantage of guidance begins to recede only when learners have sufficiently high prior knowledge to provide “internal” guidance” (Pg. 75).

.Marni’s dialog moves scaffold learning through questions and presentation of media.

Learning is scaffolded during MyST dialogs in two ways. First, MyST dialogs are designed as a sequence of “mini-dialogs” that build on each other. Each mini-dialog requires students to produce spoken responses that indicate that they understand targeted concepts. For example, concepts involved in a dialog about simple serial circuits may include mini-dialogs designed to elicit explanations in which the student indicates that they understand that a) a circuit has a specific set of components source (D-Cell), insulated wires, receiver (light bulb or motor), b) that the components have metal contact points that must touch each other to create a complete pathway, c) that electricity flows through the circuit in one direction, from the source (D-cell) through the receiver (e.g., light, motor), and back into the source, and d) electricity flows out negative side of the D-cell, through the receiver, and back into the positive side of the D-cell.

Second, within each mini-dialog, Marni’s dialog moves—her questions and the system’s presentation of media—are designed to provide the student with new information he or she can use to reason about the science and arrive at a correct answer. The system’s estimate of the students’ current state of knowledge, and the presentation of questions and media that provide the student with information, represents the process of scaffolding of learning within the students’ zone of proximal development, the zone in which the student can use new information provided by the system to build on prior knowledge to construct and share new knowledge.

For example, if the student has demonstrated that they understand that electricity flows through a circuit in one direction, but have not indicated that they understand the relationship between direction of flow and the terminals of the D-Cell, Marni will present an animation showing electricity flowing through a circuit. She will then ask: “What more can you tell me about the direction of flow?” If the student says: “I think it has something to do with the D-Cell” Marni may say: “Very good. What does the D-cell have to do with the direction of the flow?” If the student does not mention the terminals in their answer, Marni may ask: “What do the positive and negative terminals of the D-Cell have to do with the direction of flow?” If the student says

“I see that the electricity comes out of the negative side and into the positive one.” Marni may then say: “That’s right. Now what would happen to the direction of the flow of electricity if you flipped the battery?” After the student answers, Marni may say, click on the battery and tell me what’s going on.”

Marni provides immediate formative feedback throughout each dialog session. Marni gives students both implicit and explicit feedback to their answers during tutorial dialogs. Feedback is provided to students by a) modeling the use of vocabulary and scientific discourse when rephrasing the students’ previous answer, b) by providing explicit positive feedback to correct answers, and c) by providing positive reinforcement (e.g., “That was a very good explanation.”) if the student has produced a complete explanation at the end of each mini-dialog. A significant body of research has demonstrated the critical role of formative feedback in learning (Black & Wiliam, 2006; Higgins, Hartley, & Skelton, 2002).

MyST dialogs stimulate students to construct science explanations. The dialog strategies discussed above were intended to engage, stimulate, motivate and enable students to construct accurate science explanations, and achieve the satisfaction of communicating these explanations to Marni during scientific discourse. Our analysis of children’s spoken dialogs with Marni indicates that, over the course of a 15 to 20 minute dialog, students spend about as much time talking as Marni. The results of the MyST studies suggest *that all students were able to engage in conversations with Marni*. Moreover, students who scored lowest on standardized pretests of science knowledge achieved the greatest learning gains after using MyST.

Numerous studies have demonstrated that having students produce explanations during tutoring or problem solving improves learning (King, 1994; King et al., 1998; McNamara, Levinstein, & Boonthum, 2004; Palincsar & Brown, 1984; Pine & Messer, 2000). For example, Chi et al. (1989) found that having college students generate self-explanations of their understanding of physics problems improved learning. Self-explanation also improved learning about the circulatory system by eighth grade students in a controlled experiment, (Chi, DeLeeuw, Chiu, & LaVancher, 1994; Hausmann & VanLehn, 2007) . (Hausmann & VanLehn, 2007b) note that “self-explaining has consistently been shown to be effective in producing robust learning gains in the laboratory and in the classroom.” Their experiments (2007b) indicate that it is the process of actively producing explanations, rather than the accuracy of the explanations, that makes the biggest contribution to robust learning gains. MyST is all about having students construct, reflect on, refine and/or modify their explanations.

Theory and Research in Multimedia Learning

Research in multimedia learning has led to established principles for optimizing learning and enabling learners to create rich multimodal representations of science phenomena and systems. Research by Richard Mayer and colleagues has led to a vital research community (Mayer, 2001, 2003, 2005) that has established a number of principles for optimizing learning by combing spoken explanations with media. Mayer (2001) investigated students’ ability to learn how things work (motors, brakes, pumps, lightning) when information is presented in different modalities; e.g., text only, narration of the text only, text with illustrations, narrations with sequences of illustrations and narrated animations. A key finding of Mayer’s work is that simultaneously presenting spoken explanations with visual information (e.g., a sequence of illustrations or an animation) results in the highest retention of information and application of knowledge to new tasks. Mayer argues that when a person is presented with a narrated animation, the auditory and

visual modalities are processed independently and in parallel and integrated to produce an enriched mental representation. Lemke (2006, 2012) has also discussed the critical importance of multimedia in science literacy and practice.

Mayer (2001)'s cognitive theory of multimedia learning holds that well-designed narrated animations provide an optimal way to present concepts because learners construct enriched multimodal representations of knowledge that integrate verbal and visual information. Based on three assumptions—separate processing of verbal and pictorial material, limited capacity in each channel, and active construction of knowledge—Mayer five steps in his cognitive theory of multimedia learning. The learner must (1) select relevant words from the verbal input (presented as speech or text), (2) organize the words into a verbal model that makes sense of the verbal input (e.g., as a causal sequence), (3) select relevant images from pictures or animations, (4) organize the images into a pictorial model that provides a structured representation of knowledge in terms of these images, and (5) integrate word-based and image-based representations with each other and with prior knowledge to create a new mental model in long term memory.

Research on Human Tutoring

A substantial body of research has demonstrated that learning is most effective when students receive individualized instruction in small groups or one-on-one tutoring. Bloom (Bloom, 1984b) summarized studies that demonstrated that the difference between the amount and quality of learning for students who received classroom instruction relative to students who received either one-on-one or small group tutoring was up to 2 standard deviations. Evidence that tutoring works has been obtained from dozens of well-designed research studies (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001) and meta- analyses (Cohen, Kulik, & Kulik, 1982), and positive outcomes obtained in large-scale tutoring programs (Bloom, 1984a; Madden & Slavin, 1989; Topping & Whiteley, 1990).

Research on Intelligent Tutoring Systems

Research and development efforts conducted over the past two decades have resulted in Intelligent Tutoring Systems that produce learning gains equivalent to human tutoring. A recent meta-analysis by VanLehn (2011) compared learning gains achieved by students who received one-on-one tutoring with human or Intelligent Tutoring System (ITS), using stringent criteria for selection of studies based on methodological rigor. The studies included human tutoring and intelligent tutoring systems in STEM topics. When compared to students who did not receive tutoring, the effect size of human tutoring across studies was $d=0.79$ whereas the effect size of tutoring systems was $d=0.76$. VanLehn concluded that intelligent tutoring systems “are nearly as effective as human tutoring systems.” (VanLehn, 2011) (pg. 197).

VanLehn (2011) also conducted a review of the human and ITS literature to assess evidence for eight different hypotheses that have proposed to explain why tutoring is so effective in improving learning. Of the eight hypotheses considered, all but two were rejected for lack of consistent evidence. The two hypotheses validated by scientific evidence were: 1) tutoring is effective because tutors are able to *scaffold learning* by providing students with questions or hints that stimulate reasoning and enable students build on existing understandings to construct new knowledge, and 2) effective tutors provide students with *timely and meaningful feedback*, contingent on their performance. Based on these conclusions, we have focused on more effective and timely ways to scaffold learning and provide feedback to students by responding to their visual as well as their spoken behaviors.

