My Science Tutor and the MyST Corpus

Technical Report · February 2019

Ronald A. Cole
Boulder Learning Inc.
266 PUBLICATIONS **6,306** CITATIONS

Wayne Ward
University of Colorado Boulder
156 PUBLICATIONS **3,933** CITATIONS

A.

Project Overview

My Science Tutor and the MyST Corpus

Wayne Ward, Ron Cole and Sameer Pradhan

Boulder Learning Inc.

My Science Tutor: Project Overview

The 13-year **My Science Tutor (MyST) project** was conducted at Boulder Learning (formerly Boulder Language Technologies) between 2007 and 2019. It engaged over 1200 3rd, 4th and 5th grade children in spoken dialogs with a virtual science tutor. In summative evaluations, conversations with the virtual science tutor produced learning gains equivalent to gains achieved with expert human tutors, relative to students who did not receive supplemental tutoring, with moderate to strong effect sizes. The project was supported by over $9 million in research grants from the NSF and IES.

This Project Report includes published articles and a book chapter that describe the evolution and outcomes of the MyST project. The articles describe the process used to develop spoken dialog sessions, the results of the summative evaluations, and analyses of the performance of the spoken dialog system, and feedback from teachers and students. The Project Report also includes some of the proposals, reviews, and final reports submitted to the granting agencies.

To our knowledge, My Science Tutor is unique: it is the only tutoring system developed to date that supports natural spoken dialogs between a virtual tutor and children. A short video of a student interacting with the tutor can be viewed at https://www.youtube.com/watch?v=nkDmrgrVQD0&t=2s.

The content covered during MyST spoken dialog sessions was aligned to classroom science instruction using the Full Option Science System (FOSS) modules, which typically last 8 weeks during the school year. FOSS is used by over 1 million children in over 100,000 classrooms in all 50 states in the U.S. FOSS modules are centered on science investigations. There are typically 4 Investigations in a module (e.g., in the Magnetism and Electricity module, the 4 investigations are Magnetism, Series circuits, Parallel Circuits, and Electromagnetism). Each Investigation has 3 to 4 classroom "Investigation Parts" where groups of students work together to, for example, build a series circuit to make a motor run, and record their observations in science notebooks. Shortly after conducting each classroom investigation, students interacted with a virtual tutor, Marni, for 15-20 minutes. The tutor asked the student questions about science presented in illustrations, animations or interactive simulations, with follow-up questions and media designed to stimulate reasoning and help students construct accurate explanations.

MyST dialogs are strict turn-taking; the tutor presents information, asks a question about science presented in media (illustrations, animations, or interactive simulations), and waits for the student to respond. To respond, the student presses the spacebar on the laptop, holds it down while speaking, and releases it when done. Each student turn is recorded as a separate audio file. When transcribed, a session level transcript file is created for each audio file. No identifying information is stored with the data, only code numbers for schools and students.

The process of developing spoken dialogs with a virtual science tutor began with human tutors. The sessions, which were recorded and reviewed by the dialog development team, indicated the need for on-screen media—illustrations, silent animations, and interactive simulations—during tutoring sessions, so students could visualize and interact with the science they were attempting to explain in response to the tutor's questions. Dr. Margaret (Moddy) McKeown taught us how to use principles of Questioning the Author, an effective approach to classroom discourse, to pose open-ended questions to students about the science presented in the media, and how to design follow-on questions and media to scaffold learning to stimulate students to reason about the science. Once these basic principles were established, we embarked on months of Wizard of Oz studies. In these studies, human tutors in our offices at Boulder Language Technologies observed students interacting with the virtual tutor in their schools. After each student response, the human tutor would review the virtual tutor's spoken question and the media the system was about to present, and either allow the system to proceed, or type in a different response, and possibly select new media to accompany the question. Most of these sessions were reviewed and critiqued by Dr. McKeown, which helped us learn how to ask students questions to stimulate reasoning and optimize learning.

After months of reviewing these sessions, with help from Dr. McKeown, we believed we had acquired enough expertise about student responses to the tutors' questions to independently design and implement dialog sessions for student use. Basically, we imagined the various ways students could express their ideas about the specific points they needed to express to produce a correct answer to the tutor's question. (Spoiler alert: student's speech was far more creative than we imagined!)

The process of developing 16 individual 15-to-20-minute dialog sessions for each science module required a team of 7 to 8 full-time staff, led by Wayne Ward, who conceptualized and programmed the My Science Tutor system architecture. Ron Cole helped Wayne manage the project. Dani Bolanos developed and trained the Bavieca speech recognition system, Jeannine Moineau (Myatt) led the dialog development team, and Liam Devine created several thousand illustrations, animations and interactive simulations over the course of the project. (For a complete list of contributors, see authors and acknowledgements in articles and reports in Section B.)

Providing the initial release of MyST dialogs to schools for student use ***did not conclude the dialog development process***. In order to provide sufficient coverage of the diverse and creative ways students expressed their ideas in speech, it was necessary to annotate a substantial portion of the dialog sessions. Human annotators reviewed the output of the speech recognizer, and the parsing of students' utterances by the Phoenix Spoken Dialogue system. Utterances that were judged by the annotators as correct explanations, but were not identified as correct answers by the system, were then added to the grammars to improve coverage. This iterative process continued until approximately 90% of all correct explanations were parsed.

MyST Children's Speech Corpus; Project Overview

The **MyST Corpus** consists of utterances produced by students during spoken dialogs with the virtual tutor. The Corpus consists of 499 hours of conversational speech. Approximately 42% of the corpus is annotated at the word level.

The MyST Corpus *is approximately an order of magnitude larger than all other available English children's speech corpora combined.* **Boulder Learning is in the process of making the MyST Corpus freely available to the research community.** Utterances in the corpus have been partitioned into Development and Test sets to enable comparison of recognition results by the research community.

We are currently working to fully automate the process of providing the corpus to the research community. Detailed information about the Corpus is provided on the Boulder Learning web site at http://boulderlearning.com/products/corpora/. If you feel you need access to the corpus immediately for your research, please email rcole@boulderlearning.com.

The Version 1.0 release of the Corpus consists of 499 hours of children's speech collected from 1,372 third, fourth, and fifth grade students. Students conversed with the virtual science tutor Marni in 8 areas of science, resulting in a total of 11,398 student sessions and a total of 244,069 utterances. 42% of the utterances have been transcribed at the word level. If you are interested in increasing the value of the corpus to the research community, *you can contribute to the MyST Corpus by providing word-level transcriptions to utterances not yet transcribed*. Contact rcole@boulderlearning.com to contribute to the transcription effort.

Students	**Sessions**	**Total Utterances**	**Transcribed Utterances**
1,372	11,398	244,069 (499 hours)	103,429 (233 hours)

Permission to collect and distribute the student data in the MyST Corpus was approved by the University of Colorado Institutional Review Board (IRB) to assure student privacy. The IRB approved both parental consent forms and student assent forms. Parents who signed the Parental Consent form gave us permission, by checking a box on the consent form, to incorporate their child's speech into the MyST corpus, and to distribute the corpus for research or commercial use. Students also signed an assent form that they either read or had read to them, which they then signed. Student signatures were witnessed by a member of the MyST project team. Children's speech data was included in the MyST corpus if and only if both the Parental Consent and Student Assent forms were signed. The MyST corpus includes data from student sessions in multiple elementary schools in three Colorado school districts. Demographic information is reported in the published journal articles.

Organization of the Project Report

Section B describes the MyST Children's Speech Corpus. It compares the MyST Corpus to other available English children's speech corpora in terms of the amount and type of speech data collected (e.g., prompted versus spontaneous speech), and the cost of obtaining each corpus for research or commercial use.

Section C provides detailed information about the MyST Project. The Project Report includes published articles, a book chapter, the final report of the IES Innovation and Development grant, the final report of the NSF Discovery Research K-12 grant, and results of the Replication study conducted in 2015 during the first year of the IES Replication and Efficacy study. (Section B will be updated with the results of the Efficacy study when the final report is completed.)

B.

The MyST Children's Speech Corpus

We are featuring the MyST Corpus as the first section of the My Science Tutor project report because a) we are eager to announce its immediate availability to the research community, and b) we hope the corpus will accelerate research and development of children's speech recognition systems with great potential benefits, e.g., development of Intelligent Tutoring Systems that support spoken dialogs with virtual agents, or using speech recognition to assess and facilitate children's reading comprehension.

Section B.1 provides a summary of the MyST Corpus. Additional information about the corpus, and how to get it, can be found on the Boulder Learning Web site.
http://boulderlearning.com/products/corpora/.

Section B.2 estimates the size and features of the MyST Corpus, relative to other available English Children's speech corpora. In terms of *spontaneous speech*, the MyST Corpus has about an order of magnitude more children's speech data than all other children's speech corpora combined.

If you download the MyST corpus, please help us by providing feedback on any problems encountered.

If you would like to help make the corpus a more valuable resource, please transcribing portions of the corpus that have not yet been transcribed at the word level. About 55% of the corpus has yet to be transcribed. We have learned that transcribing children's spoken dialogs is a great learning experience.

If there is enough interest in contributing to the corpus, we will provide sets of utterances and transcription conventions to interested researchers. Ideally, sets of utterances could be transcribed by independent transcribers, so recognizers could be trained on words identified as spoken correctly by independent transcribers.

After the MyST Corpus has been vetted by the research community we plan to make the Corpus available for commercial use for a fee. In section 2, we made best efforts to identify the cost of all available English children's speech corpora available for commercial use, so we can set a fair price for the corpus for commercial use.

B.1

The MyST Children’s Speech Corpus (V 1.0)

The version 1.0 release of the MyST Corpus consists of 499 hours of children's speech collected from 1,372 third, fourth, and fifth grade students. Students conversed with a virtual science tutor in 8 areas of science, resulting in a total of 11,398 student sessions and a total of 244,069 utterances. 42% of the utterances have been transcribed at the word level.

Students	Sessions	Total Utterances	Transcribed Utterances
1,372	11,398	244,069 (499 hours)	103,429 (233 hours)

Data Collection

The MyST corpus was collected in 2 stages, Phase I and Phase II

In both phases, the content covered is aligned to Full Option Science System (FOSS) modules, which typically last 8 weeks during the school year. FOSS is used by over 1 million children in over 100,000 classrooms in all 50 states in the U.S. FOSS modules are centered on science investigations. There are typically 4 Investigations in a module (e.g., in the Magnetism and Electricity module, the 4 investigations are Magnetism, Serial circuits, Parallel Circuits, and Electromagnetism). Each Investigation has 3 to 4 classroom “investigation parts” where groups of students work together to, for example, build a serial circuit to make a motor run, and record their observations in science notebooks. Shortly after conducting an investigation, students interact with the virtual tutor for 15-20 minutes. The tutor asks the student questions about science presented in illustrations, animations or interactive simulations, with follow-up questions designed to stimulate reasoning and help students construct accurate explanations.

The system is strict turn-taking; the tutor presents information, asks a question and waits for the student to respond. To respond, the student presses the spacebar on the laptop, holds it down while speaking, and releases it when done. Each student turn is recorded as a separate audio file. When transcribed, a session level transcript file *trs is created for each audio file. No identifying information is stored with the data, only code numbers for schools and students. All students and their parents signed consent forms allowing Boulder Learning to enter and distribute their anonymous speech data.

The file structure for both corpora is

corpora/myst/data/<student_id>/<session_id>/<session_id>.<file-extension>

Phase I

The Phase I corpus contains sessions from students in grades 3-5. All of the sessions have been transcribed.

1. ME - Magnetism and Electricity
2. MS - Mixtures and Solutions
3. VB - Variables
4. WA - Water

A <session_id>is represented as
<corpus>_<student_id>_<date>_<time>_<module>_<investigation>.<part>

The <student_id>is a 3-digit school code and a 3-digit student number. example:
myst_990507_2010-04-02_00-00-00_WA_2.3
We did not capture the time for Phase I, so all the times are 00-00-00

Number of Students:	421
Number of Sessions:	1512 (110 hours)
Transcribed Sessions:	1512 (110 hours)

There was no attempt to have any individual student cover all of the parts for a module. The focus of the collection was to get a wide variety of students rather than try to get complete coverage of material for individual students.

Phase II

The Phase II corpus contains sessions from students in grades 4-5. It uses 5 modules, with an average of 10 parts each

1. EE - Energy and Electromagnetism
2. MX - Mixtures
3. SMP - Sun, Moon and Planets
4. SRL - Soil, Rocks and Landforms
5. LS - Living Systems

Again, the student_id is a 3-digit school code and a 3-digit student number. The session ids are encoded similarly.

Number of Students:	951		
Number of Sessions:	9,886	(389	hours)
Transcribed Sessions:	1,426	(112	hours)
Untranscribed Sessions:	3,711	(277	hours)

B.2

Valuing the MyST Corpus

Comparing the MyST Corpus to other Kids' Speech Corpora

Overview

In this section, we compare MyST Corpus to all other English Kids' Speech Corpora. Information about each corpus was obtained from the offerors' Web sites, or the Linguistic Data Consortium (https://www.ldc.upenn.edu/). We made best efforts to determine the amount of children's speech data collected, the type of data collected (e.g., prompted versus conversational speech), and the cost of the corpus for research and commercial use. When information was difficult to find, we attempted to contact the appropriate individuals in the organizations that distribute the corpus.

Short Answer*: The MyST Corpus contains about 5 times more speech than all other available English children's speech corpora combined.* "Other available children's speech corpora," described below, contain prompted speech (individual words, digit strings, prompted utterances), and/or children reading aloud. *The MyST Corpus consists solely of speech collected from children during conversations with a virtual tutor.* **In terms of conversational speech, *The MyST Corpus has more than an order of magnitude more conversational speech than all other children's speech corpora combined.***

An article published by Bolanos et al. 2011 provides a credible insight into the total amount of children's English speech data available to researchers in 2011. The article describes FLORA, a speech verification system that used the Bavieca recognizer to score grade-level texts read aloud by elementary students. The recognizer was trained on all available data from the OGI Kids' Speech Corpus, and the CU Kids' Speech Corpus. Here is the description:

"*Training corpora.* Three different speech corpora were used to train the acoustic models used by FLORA. The University of Colorado Read and Summarized Stories Corpus [Cole and Pellom 2006] (325 speakers from 1st to 5th grade), the CU Read and Prompted Children's Corpus [Cole et al. 2006] (663 speakers from Kindergarten through 5th grade) and the OGI Kids' Speech Corpus [Shobaki et al., 2000] (509 speakers from 1st to 5th grade). A total of 106 hours of speech from these corpora was used to train the acoustic models." [Bolaños, D., Cole, R. A., Ward, W., Borts, E., and Svirsky, E. (2011). FLORA: Fluent oral reading assessment of children's speech. ACM Trans. Speech Lang. Process. 7, 4, Article 16.]

Note: *Development of both the OGI Kids' Speech Corpus and the CU Kids' Speech Corpus were conceptualized and supervised by Ron Cole, one of the authors of this project report. The OGI*

Kids' Speech Corpus was developed while Ron was Director of the Center for Spoken Language Understanding (CSLU) at the Oregon Graduate Institute in collaboration with Kal Shobaki and Paul Hosom. The CU Kids' Speech Corpus was developed when Ron was Director of the Center for Spoken Language Research (CSLR) at the University of Colorado. Both corpora were freely distributed to the research community.

Other Available Children's Speech Corpora

There are **3 available English children's speech corpora** that can be obtained from the Linguistic Data Consortium or directly from the institutions that own them.

These corpora are:

1. **The CMU Kids' Speech Corpus: ($500 for Non-Members of the Linguistic Data Consortium).**

Information about the CMU Kids' speech corpus can be found at the Linguistic Data Consortium at https://catalog.ldc.upenn.edu/LDC97S63. The corpus consists of approximately 9 hours of speech read aloud by 24 male and 52 female speakers in first, second and third grades. The corpus consists of 5,180 utterances. We estimate, based on the number of sentences read aloud, that the corpus consists of approximately **9 hours of read speech**.

2. **The OGI Kids' Speech Corpus ($5500 from Oregon Health & Sciences University; $150 from Linguistic Data Consortium).**

Information about the OGI Kids' Speech Corpus can be found at the Oregon Health & Science University (CORPORA from CSLU: Kids speech v1.1.
https://ibridgenetwork.org/#!/profiles/1082945057800/innovations/31/

The CSLU Kids' Speech Version 1.1 is also available from the Linguistic Data Consortium for **$150 (a $5,350 savings compared to the OHSU price)**
https://catalog.ldc.upenn.edu/LDC2007S18

The corpus consists of prompted and spontaneous speech from 1100 from children in kindergarten through 10th grade. Prompted speech read aloud by each student consisted of individual words, digit strings and sentences. In addition, each student produced spontaneous speech in answer to an open-ended question. The total corpus comprises approximately **150 total hours of speech**. Approximately, **and ~17 hours of spontaneous speech.**

Detailed information about the CSLU Kids' Speech corpus can be found in:
Shobaki, K., Hosom, J-P., & Cole, RA. (2000). The OGI Kids' Speech Corpus and Recognizers. Proceedings of the Sixth International Conference on Spoken Language Understanding (ICSLP). Beijing, China, October 16-20, 2000. https://www.isca-speech.org/archive/archive_papers/icslp_2000/i00_4258.pdf

3. The CU Kids' Speech Corpus ($10,000 CU Boulder Tech Transfer Office).

Information about the CU Kids' Speech Corpus can be obtained from the University of Colorado Boulder Technology Transfer Office. (A Google search of CU Kids' Speech Corpus produces a PDF at https://content.cu.edu/bouldertech/show_NCSum.cfm?NCS=1511460. However, the PDF is out of date; Kate Tallman, the point of contact, no longer works at the CU Boulder Tech Transfer Office to obtain the corpus.)

The cost of the commercial license for the Corpus, $10,000, is based on a 2002 license agreement I that includes the following paragraph and table:

6.2 License Fee Amount. Payment. The license fee for all CSLR corpora for companies shall be a one-time fee of $25,000 (twenty-five thousand dollars), payable upon execution of this Agreement. Payment will be sent to University of Colorado, Suite 390, 589 UCB, Boulder, CO U.S.A. 80309-0589. Individual corpora can be purchased individually according to the prices in the Table below:

Corpus	Not for profit U.S. organizations	Not for profit foreign organizations	For Profit organizations
Kid's speech	$500	$1500	$10,000
CU move	$2,500	$5000	$15,000
Communicator	$200	$1000	$10,000

Please contact CU Tech Transfer Office if you are interested in these corpora. The costs displayed in this table may be out of date.

Two technical reports were published about the CU Kids' speech corpus:

Cole, R., Hosom, P., and Pellom, B. 2006. University of Colorado prompted and read children's speech corpus. Tech. rep. TR-CSLR-2006-02, Center for Spoken Language Research, University of Colorado, Boulder.

Cole, R. & Pellom, B. 2006. University of Colorado read and summarized stories corpus. Tech. rep. TRCSLR-2006-03, Center for Spoken Language Research, University of Colorado, Boulder.

Unfortunately, we have been unable to locate either report.

However, the following article provides a detailed description of the CU Kids' Speech Corpus:

Hagen, A., Pellom, B., & Cole, R. "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Communication* 49; 861–873. Here is what's reported on pages 864 – 866.

The article can be found at:
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.911.3864&rep=rep1&type=pdf

We have printed the relevant section here:

7. CU Children's audio speech corpora

The research conducted in this paper makes use of two new children's speech corpora collected by (Cole and Pellom, 2006a, 2006a, 2006b) at the University of Colorado.

7.1 University of Colorado Prompted and Read Children's speech corpus

The CSLR Prompted and Read Children's speech corpus consists of transcribed speech data collected from 663 Kindergarten through fifth grade children producing isolated words, sentences, and short spontaneous stories. The protocol is described in (Cole and Pellom, 2006a). Table 1 provides the number of speakers per grade level.

Each speaker produced approximately 100 utterances which vary in length depending on the protocol. The recordings were made using one of three types of micro- phones: a commonly available head-mounted noise-canceling microphone (Labtec LVA-8450), an array microphone (CNnetcom-Voice Array Microphone VA-2000), and a commonly available desktop far field microphone. The final corpus is sampled at 16 kHz at 16 bits per sample. Each audio file is accompanied by a word-level transcription. Corresponding information such as subject ID, age, sex, grade-level, and native language of speaker is also provided.

7.2 University of Colorado Read and Summarized Stories Corpus

The CU Read and Summarized Stories Corpus (Cole and Pellom, 2006b) consists of transcribed speech data from 106 children in grades 3-5 within the Boulder Valley School District (Grade 3: 17 speakers, Grade 4: 28 speakers, Grade 5: 61 speakers) who read and summarized stories during a 30 min session from a set of 10 stories, and 221 children in grades 1 and 2 who summarized stories read to them from a set of 62 stories.

Third through fifth grade children also read 25 phonetically balanced sentences for future use in exploring strategies for speaker adaptation. Data were collected in a quiet room using a Labtec Axis-502 microphone. The data were recorded at 44 kHz and later re-sampled to 16 kHz for the purposes of experimentation. The current corpus consists of 10 different stories. The number of speakers per story is shown in Table 2. Each story contained an average of 1054 words (min 532 words/max 1926 words) with an average of 413 unique words per story. The resulting summaries spoken by children contain an average of 168 words. The additional children are used for acoustic model training in our current system together with the CU Read and Prompted Children's Corpus and the OGI Kids' Speech Corpus (Shobaki et al., 2000).

Table 1
Subject count in the CU Prompted and Read Children's speech corpus by grade level

Grade	*K*	*1*	*2*	*3*	*4*	*5*
#Children	*84*	*136*	*150*	*92*	*91*	*110*

Refs

Cole, R., Hosom, P., Pellom, B., 2006a. University of Colorado Prompted and Read Children's Speech Corpus. Technical Report TR-CSLR-2006-02, Center for Spoken Language Research, University of Colorado, Boulder.

Cole, R., Pellom, B., 2006b. University of Colorado Read and Summarized Stories Corpus. Technical Report TR- CSLR-2006-03, Center for Spoken Language Research, University of Colorado.

Shobaki, K. Hosom, J.P., Cole, R. 2000. The OGI Kids' Speech Corpus and Recognizers. In: Proceedings of ICSLP 2000, Beijing, China.

C.

My Science Tutor

2007 – 2019

Project Overview

My Science Tutor is an Intelligent Tutoring System designed to engage 3rd, 4th and 5th grade students in spoken dialogs with a virtual tutor. Over the course of the project, students interacted with the virtual tutor Marni in 8 areas of science. The science modules were part of the Full Option Science System (FOSS), an inquiry-based program aligned to US National Science Standards. Depending upon the science module, students interacted with Marni for 15 to 20 minutes in 8 to 16 sessions.

Dialog sessions occurred soon after students conducted classroom science investigations in small groups. The content of each tutoring session was aligned to the science concepts and phenomena students encountered in their science investigations and wrote about in their science notebooks. During each tutoring session, Marni asked open-ended questions about the science presented in illustrations, silent animations of interactive simulations, and students produced spoken answers to her questions.

Three independent summative evaluations were conducted that compared learning gains of students who were tutored by the virtual tutor Marni or were tutored by expert human tutors. Marni and the human tutors used similar "dialog moves" based on an effective approach to classroom discourse (Questioning the Author) and had access to the same set of media. The summative evaluations compared pre-test versus post-test learning gains using valid and reliable FOSS assessments. Results of all three assessments indicated statistically equivalent moderate to strong learning gains by human tutors and by Marni, compared to students who did not receive supplemental tutoring.

The MyST project was funded by grants of approximately ~$3 million each from:

- The NSF Discovery Research K-12 (DRK-12) program in the Education and Human Resources (EHR) Directorate
- An IES Innovation and Development grant from the Cognition and Student Learning (CASL) Program
- An IES Replication and Efficacy grant from CASL program to conduct a Replication and Efficacy study.

The MyST Project Report includes 3 published articles, a book chapter, and the Final Project Reports submitted to the NSF DRK-12 and the IES CASL program directors. The results of the IES Replication study, conducted during the first year of the IES Replication and Efficacy project, will be included in this Project Report as soon as it is available.

C.1

My Science Tutor Publications

The published articles below describe the scientific rationale for MyST dialogs, the process used to develop spoken dialog sessions, the results of 3 summative evaluations, and analyses of the performance of the spoken dialog system. Three independent studies compared learning gains of students who were either tutored by Marni, or by expert human tutors, compared to students who did not receive tutoring. The results showed equivalent learning gains for students tutored by Marni or by Human tutors, with moderate to strong effect sizes. To our knowledge, MyST is the only intelligent tutoring system that supports natural spoken dialogs with Children in multiple areas of science.

MyST Publications

Journal Articles

Cole, R., Buchenroth-Martin, C., Weston, T., Devine, L., Myatt, J., Helding, B., Pradhan, S., McKeown, M., Messier, S., Borum, J., & Ward, W. (2018). One-on-one and small group conversations with an intelligentvirtual science tutor. *Computer Speech and Language*, 50, 157-174.

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., & Weston, T. (2013). My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology, (Special Issue on Advanced Learning Technologies)*. 105(4), 1115-1125.

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., & Vuuren, S. V. (2011). My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Transactions on Speech and Language Processing (TSLP), Special Issue on Speech and Language Processing of Children's Speech for Child-machine Interaction, 7(4).*

Book Chapter

Ward, W. & Cole R. (2015). Developing Conversational Multimedia Tutorial Dialogs. Design Recommendations for Intelligent Tutoring Systems. Volume 3: Authoring Tools & Expert Modeling Techniques. Edited by Robert Sottilare, Arthur Graesser, Xiangen Hu, Keith Brawne. Chapter 20, pp. 243 – 254. Published by US Army Research Laboratory.

Conference Proceedings

Wayne Ward, Daniel Bolaños, Ronald A. Cole: Spoken Dialogs with a Virtual Science Tutor, INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012.

Sameer Pradhan, Ronald A. Cole, Wayne Ward: My Science Tutorial—Learning Science with a Conversational Virtual Tutor, ACL 2016, Proceedings of 54th Annual Meeting of the Association of Computational Linguistics—System Demonstrations, Berlin, Germany, August 7-12, 2016

My Science Tutor: A Conversational Multimedia Virtual Tutor for Elementary School Science

WAYNE WARD, RONALD COLE, DANIEL BOLAÑOS, CINDY BUCHENROTH-MARTIN,
and EDWARD SVIRSKY, Boulder Language Technologies
SAREL VAN VUUREN, TIMOTHY WESTON, JING ZHENG,
and LEE BECKER, University of Colorado, Boulder

This article describes My Science Tutor (MyST), an intelligent tutoring system designed to improvescience learning by students in 3rd, 4th, and 5th grades (7 to 11 years old) through conversational dialogs with a virtual science tutor. In our study, individual students engage in spoken dialogs with the virtual tutor Marni during 15 to 20 minute sessions following classroom science investigations to discuss and extend concepts embedded in the investigations. The spoken dialogs in MyST are designed to scaffold learning by presenting open-ended questions accompanied by illustrations or animations related to the classroom investigations and the science concepts being learned. The focus of the interactions is to elicit self-expression from students. To this end, Marni applies some of the principles of *Questioning the Author*, a proven approach to classroom conversations, to challenge students to think about and integrate new concepts with prior knowledge to construct enriched mental models that can be used to explain and predict scientific phenomena. In this article, we describe how spoken dialogs using Automatic Speech Recognition (ASR) and natural language processing were developed to stimulate students' thinking, reasoning and self explanations. We describe the MyST system architecture and Wizard of Oz procedure that was used to collect data from tutorial sessions with elementary school students. Using data collected with the procedure, we present evaluations of the ASR and semantic parsing components. A formal evaluation of learning gains resulting from system use is currently being conducted. This paper presents survey results of teachers' and children's impressions of MyST.

Categories and Subject Descriptors: I 2.7 [**Artificial Intelligence**]: Natural language processing—*Speech recognition and synthesis*

General Terms: Design

Additional Key Words and Phrases: Semantic parsing, language model, dialog management, avatar

ACM Reference Format:
Ward, W., Cole, R., Bolañ os, D., Buchenroth-Martin, C., Svirsky, E., Van Vuuren, S., Weston, T., Zheng, J., and Becker, L. 2011. My Science Tutor: A conversational multimedia tutor for elementary school students. ACM Trans. Speech Lang. Process. 7, 4, Article 18 (August 2011), 29 pages.
DOI = 10.1145/1998384.1998392 http://doi.acm.org/10.1145/1998384.1998392

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B070008, and the National Science Foundation through Grants DRL 0733323 to Boulder Language Technologies and DRL 0733322 to the University of Colorado at Boulder. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education or the National Science Foundation.

W. Ward is Chief Scientist and shareholder at Boulder Language Technologies and a research professor at the University of Colorado, Boulder
Author's address: email: wward@bltek.com

DOI 10.1145/1998384.1998392 http://doi.acm.org/10.1145/1998384.1998392

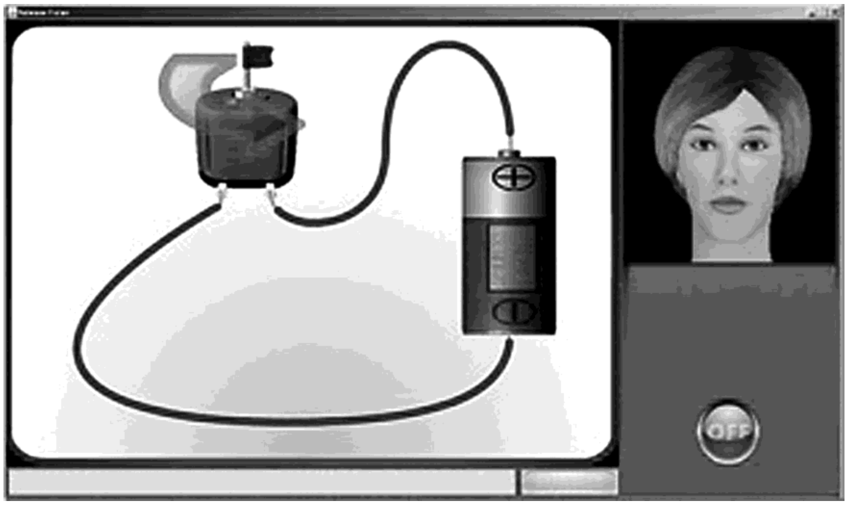


Fig. 1. Virtual Tutor Screen.

1. INTRODUCTION

There is a clear and urgent need to develop accessible and effective learning tools to supplement and improve classroom science instruction for many students in the United States. According to the 2005 National Assessment of Educational Progress (NAEP 2005) only three percent of U.S. students attained advanced levels of science achievement in Grades 4 and 8 and only two percent reached advanced levels in Grade 12.

Since 2007, our research team has been involved in an intensive effort 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 requires the integration of 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 displays a screenshot of the virtual tutor Marni asking questions about media displayed in tutorial dialog sessions.

Over the past decade, advances in speech and language processing have enabled research and development of a growing number of intelligent tutoring systems that use spoken dialogs to tutor children and adults [Mostow and Aist 2001; Rickel and Johnson 2000; Graesser et al. 2001; Aist and Mostow 2009; Mostow and Chen 2009; Chen et al. 2010]. These systems have focused mainly on science, reading, and language learning. Our literature review indicated that science tutors that incorporate spoken dialogs have been designed for use by university-level students [Graesser et al. 2001; Littman and Stillman 2004]. Science tutors have been developed for children that incorporate embodied conversational agents (computer character that talk) in multimedia environments [Lester et al. 1997, 1999; Dede et al. 2010], but these systems do not support natural spoken dialogs between a child and the agent. Spoken dialogues with children have been used successfully to help children learn to read and comprehend text and to assess an individual's proficiency in a given language. For example, work in Project Listen integrated speech recognition and dialogue modeling to improve reading, vocabulary acquisition and text comprehension [Aist and Mostow 2009; Mostow and Chen 2009; Chen et al. 2010]. Bernstein and Cheng [2007] demonstrated the validity of scores from fully automated tests that use ASR to assess a child's ability to understand and communicate in English. While spoken dialogue systems have been developed for science tutoring for university-level students, and for children for reading and language assessment, we have no evidence of intelligent tutoring systems that support spoken conversational interaction between children and a virtual science tutor. To our knowledge, MyST is unique in this regard.

The goal of the MyST project is to help struggling students learn the science concepts encountered in classroom science instruction. Each 15 to 20 minute MyST dialogue session functions as an independent learning activity that provides, to the extent possible, the scaffolding required to stimulate students to think, reason and talk about science during spoken dialogues with the virtual tutor Marni. The goal of these *multimedia dialogues* is to help students construct and generate explanations that express their ideas. The dialogues are designed so that over the course of the conversation with Marni, the student is able to 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 dialogues are linked to the activities, observations and outcomes of classroom science investigations conducted by groups of three to five children in kit-based science investigations that are part of the FOSS (Full Option Science System) program used by over one million students in classrooms in all fifty states in the United States.[1] In addition to the science kits that support an average of 16 hour-long investigations in each FOSS module (i.e., a specific area of science), the program includes a Teacher Guide (professional development for teachers on how to use the FOSS program to best effect, including helping students organize their predictions, observations and conclusions in science notebooks), a set of science stories that students may read, and valid and reliable standardized Assessments of Science Knowledge (ASK) administered to each student before after each eight to ten week module.

Within a given FOSS module, the initial investigations provide the foundational knowledge for conducting more sophisticated investigations. For example, investigations of magnetism and simple circuits lead to investigations in which children build both serial and parallel circuits, followed by investigations in which they build electromagnets and explore electromagnetism. In our study, we developed 16 different tutorial dialogue sessions, lasting about 20 minutes each, for four different areas of science: Variables, Measurement, Water and Magnetism and Electricity. Thus, a total of 64 different tutorials, were developed across the four areas of science to help children think about and explain science concepts encountered during classroom activities.

Conversations with Marni are characterized by two key features: the inclusion of media, in the form of an illustration, animation or interactive simulation throughout the dialogue, and the use of open-ended questions related to the phenomena and concepts presented via the media. For example, an initial classroom investigation about magnets has students move around the classroom exploring and writing down what things do and do not stick to their magnets. The subsequent multimedia dialogue with Marni begins with an animation that shows a magnet being moved over a set of identifiable objects, which picks up some of the objects but not others. Marni then says, "What's going on here?" If the student says, "The magnet picked up some of the objects," Marni might say, "Tell me more about that." To use another simple example, following a classroom investigation about circuits in which children work together to build a circuit using a battery, wires, a switch and a light bulb, the tutorial begins a picture of the circuit components, with Marni asking, "What's this all about?"

[1] www.fossweb.com.

In the remainder of this article, we present the scientific rationale for MyST, describe the system architecture and technologies that support conversations about science with Marni in multimedia environments, and describe the development of a corpus of conversational tutorial sessions. Using the corpus, we present evaluations of the ASR and semantic parsing and dialogue components of the system. In addition to component level evaluations, the MyST project will also assess the system along the dimensions of Engagement (how satisfactory is the user experience?), feasibility (can the system be used in the way proposed in real world situations?), and efficacy (does the system produce learning gains?). A formal evaluation of these aspects of the system is currently being conducted. While these data are not yet available, this paper presents survey results of teachers' and children's impressions of MyST from the data collection done in the 2009–2010 academic year. These surveys give evidence for the Engagement and Feasibility of the system.

2. SCIENTIFIC RATIONALE

MyST is an example of a new generation of intelligent tutoring systems that facilitate learning through spoken dialogues with a virtual tutor in multimedia activities. Intelligent tutoring systems aim to enhance learning achievement by providing students with individualized instruction similar to that provided by a knowledgeable human tutor. These systems support typed or spoken input with the system presenting prompts and feedback via text, a human voice, or an animated pedagogical agent [Graesser et al. 2001; Wise et al. 2005; Lester et al. 1997; Mostow and Aist 2001]. Text, illustrations, and animations may be incorporated into the dialogues. Research studies show up to one sigma gains (approximately equivalent to an improvement of one letter grade) when comparing performance of high school and college students who use the tutoring systems to students who receive classroom instruction on the same content [Graesser et al. 2001; Van Lehn and Graesser 200l; Van Lehn et al. 2005].

The development of MyST is informed by several decades of research in psychology and computer science. In the remainder of this section we describe theory and research that informed the design of MyST.

Social Constructivism. The work of Jean Piaget, Lev Vygotsky, and Jerome Bruner gave rise to a theory of cognitive development and knowledge acquisition known as social constructivism, which provides a strong rationale for the use of tutorial dialogs to optimize learning. In social constructivism, learning is viewed as an active social process of constructing knowledge "that occurs through processes of interaction, negotiation, and collaboration" [Palincsar 1998]. Vygotsky [1978] stressed the critical role of social interaction within one's culture in acquiring the social and linguistic tools that are the basis of knowledge acquisition. "Learning awakens a variety of internal developmental processes that are able to operate only when the child is interacting with people in his environment" [Vygotsky 1978]. He stressed the importance of having students learn by presenting problems that enable them to scaffold existing knowledge to acquire new knowledge. Vygotsky introduced the concept of the Zone of Proximal Development, "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]. Social constructivism provides the conceptual model for knowledge acquisition in MyST: to improve learning by scaffolding conversations using media to support hypothesis generation and coconstruction of knowledge.

Discourse Comprehension Theory. Cognitive learning theorists generally agree that learning occurs most effectively when students are actively engaged in critical thinking and reasoning processes that cause new information to be integrated with prior knowledge. Discourse Comprehension Theory [Kintsch 1988; 1998] provides a strong theoretical framework for asking questions and designing activities that stimulate thinking and construction of deep knowledge that is useful and transferable. This theory provides the foundation for several instructional approaches to comprehension [King 1991; Beck et al. 1996; Beck and McKeown 2006]. Comprehension theory holds that deep learning requires integration of prior knowledge with new information and results in the ability to use this information constructively in new contexts.

Benefits of Tutorial Instruction. Theory and research provide strong guidelines for designing effective tutoring dialogs. Over two decades of research have demonstrated that learning is most effective when students receive individualized instruction in small groups or one-on-one tutoring. Bloom [1984] determined that the difference between the amount and quality of learning for students who received classroom instruction and those who received either one-on-one or small group tutoring was 2 standard deviations. Evidence that tutoring works has been obtained from dozens of well designed research studies, meta-analyses of research studies [Cohen et al. 1982], and positive outcomes obtained in large-scale tutoring programs [Topping and Whitley 1990; Madden and Slavin 1989]. Benefits of tutoring can be attributed to several factors, of which the following three appear to contribute most.

(1) *Question generation.* A significant body of research shows that learning improves when teachers and students ask deep-level-reasoning questions [Bloom 1956]. Asking authentic questions leads to improved comprehension, learning, and retention of texts and lectures by college students [Craig et al. 2000; Driscoll et al. 2003; King 1989] and school children [King 1994; King et al. 1998; Palincsar and Brown 1984]. Nystrand and Gamarond [1991] found that genuine dialogs, although rare in the classrooms studied, were most often initiated by authentic questions asked by students.

(2) *Self explanation.* Research has demonstrated that having students produce explanations improves learning [King 1994; King et al. 1998; Palincsar and Brown 1984; Chi et al. 1989, 2001]. In a series of studies, Chi et al. [1989, 2001] 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 et al. 1994]. Hausmann and Van Lehn [2007a] note that: “self-explaining has consistently been shown to be effective in producing robust learning gains in the laboratory and in the classroom.” Experiments by Hausmann and Van Lehn [2007b] indicate that it is the process of actively producing explanations, rather than the accuracy of the explanations, that makes the biggest contribution to learning.

(3) *Knowledge coconstruction.* Students coconstruct knowledge when they are provided the opportunity to express their ideas, and to evaluate their thoughts in terms of ideas presented by others. There is compelling evidence that engaging students in meaningful conversations improves learning [Chi et al. 1989; King 1994; 1998; Palincsar and Brown 1984; Pine and Messer 2000; Butcher 2006; Soter et al. 2008; Murphy et al. 2009]. Classroom conversations and tutorial dialogs increase the opportunity for occurrences of knowledge coconstruction which has been shown to have a significant impact on learning gains [Wood and Middleton 1975; King 1994; Chapin et al. 2003; Chi et al. 2001].

Benefits of Social Agency and Pedagogical Agents. When human computer interfaces are consistent with the social conventions that guide our daily interactions with other people, they provide more engaging, satisfying, and effective user experiences [Reeves and Nass 1996; Nass and Brave 2005]. Such programs foster social agency, enabling users to interact with them the way they interact with people. In comparisons of programs with and without talking heads or human voices, children learned more and reported more satisfaction using programs that incorporated virtual humans [Moreno et al. 2001; Atkinson 2002; Baylor et al. 2005]. A number of researchers have observed that children become highly engaged with virtual tutors and appear to interact with a virtual tutor as if it were a real teacher and appear motivated to work hard to please the virtual tutor. Lester et al. [1997] termed this phenomenon the "Persona Effect." In our previous research using Marni as a virtual reading tutor [Cole et al. 2007], over 70% of over 250 students surveyed reported that they trusted Marni, that they felt Marni cared about them, that Marni was a good teacher, and that Marni helped them learn to read.

Benefits of Multimedia Presentations. The design of the proposed tutorials is informed by research on multimedia learning conducted by Richard Mayer and his colleagues (See Mayer [2001] for a review). Mayer and his colleagues investigated students' ability to learn how things work (motors, brakes, pumps, lightning) when information was presented in different modalities; for instance, text only, narration of the text only, text with illustrations, narrations with sequences of illustrations, or narrated animations. A key finding of Mayer's work is that simultaneously presenting speech (narration) 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 in a narrated animation, a student's auditory and visual modalities are processed independently but are integrated to produce an enriched mental representation.

3. MULTIMEDIA DIALOGS

Students learn science in MyST through natural spoken dialogs with the virtual tutor Marni, a lifelike 3-D computer character that is "on screen" at all times. In general, Marni asks students open-ended questions related to illustrations or animations displayed on the computer screen. The spoken dialogue system processes the student's speech to assess the student's 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 thinking and reasoning that can lead to accurate explanations, as described below. We call these conversations with Marni *multimedia dialogues*, since students simultaneously listen to and think about Marni's questions while viewing illustrations and animations or interacting with a simulation.

Marni produces accurate movements of the lips and tongue in synchrony with either recorded or synthetically generated speech. Marni's visual speech is produced fully automatically by the CU Animate system [Cole et al. 2003; Ma et al. 2004] from an input text string and acoustic waveform of the spoken words in the text string. During the initial development and refinement of the MyST system we used high quality text-to-speech (TTS) synthesis rather than recorded speech. Since dialogues were constantly evolving, it was far more efficient and cost effective to use text-to-speech synthesis rather than record new utterances each time we changed the dialogue. In addition, using TTS allowed human tutors to type in the text they wanted Marni to speak in real time while students were conversing with Marni. This type of interaction is called a Wizard of Oz procedure and is described in the following. At the conclusion of the development phase of each module, a human tutor recorded each of the prompts produced by Marni, enabling her to speak with a human voice that produced appropriate emotional expression, such as enthusiasm when reinforcing the student for accurate and complete explanations.

3.1. Questioning the Author Approach to Tutorial Dialogs

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 [Beck et al. 1996; McKeown and Beck 1999; McKeown et al. 1999].

QtA is a mature, scientifically-based and effective program used by hundred 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. The program has well established procedures for training teachers to interact with students, for observing teachers in classrooms and for providing feedback to teachers. In recent studies [Murphy and Edwards 2005; Murphy et al. 2009], QtA was identified as one of two approaches out of the nine examined that are likely to promote high-level thinking and comprehension of text. Relative to control conditions, QtA showed effect sizes of .63 on measures of text comprehension, and of 2.5 on researcher-developed measures of critical thinking/reasoning [Murphy and Edwards 2005]. Moreover, analysis of QtA discourse showed a relatively high incidence of authentic questions, uptake, and teacher questions that promote high-level thinking—all indicators of productive discussions likely to promote learning and comprehension of text [Nystrand and Gamoran 1991; Soter and Rudge 2005; Soter et al. 2008].

Questioning the Author is a deceptively simple approach. Its focus is to have students grapple with, and reflect on, what an author is trying to say in order to build a representation from it. Because the dialog modeling used in QtA is well understood, can be taught to others [Beck and McKeown 2006], and has been demonstrated to be effective in improving comprehension of informational texts, we decided to incorporate principles of QtA into tutorial dialogues within 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 dialogues, 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 iterative process of testing and refining these QtA-based multimedia dialogues.

We note that, in the context of an inquiry-based science program, the perspective of the “author” in “Questioning the Author” moves from questions about what a specific author is trying to communicate, to questions about science investigations and outcomes. In a sense, in a science investigation the “author” is Mother Nature, and the “texts” are the observations that students make and the datasets they enter into their science notebooks. During multimedia dialogues, 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.

3.2. Use of Media in MyST Dialogs

MyST dialogs typically incorporate one of three types of media 1) static illustrations, 2) simple animations and 3) interactive investigations. Although they sometimes overlap in the content presented, each media type plays a unique and important role in science learning in MyST dialogs.

Static Illustrations. Static Illustrations are inanimate Flash drawings. We have found that Static Illustrations are a good way to initiate discussions about topics. They provide the student with 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?”

The sequence of questions presented by the virtual tutor starts 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 the following.

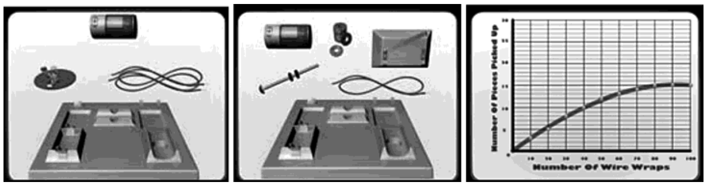
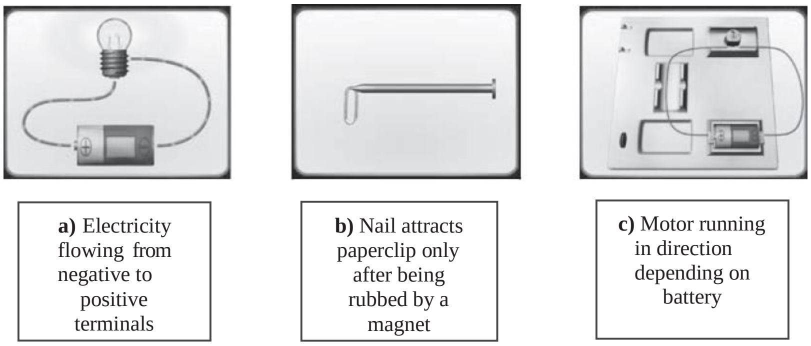


Fig. 2. Example static illustrations.



a) Electricity flowing from negative to positive terminals

b) Nail attracts paperclip only after being rubbed by a magnet

c) Motor running in direction depending on battery

Fig. 3. Example animations.

—*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 who grasps what they are looking at, but not how to interpret it. A QtA inspired sequence about the graph might be the following.

—T: What do you think this is about?
—S: I think it's a graph of something.
—T: Good observation. It is a graph of something. 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 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 noninteractive Flash animations. Simple Animations can provide additional information and 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 creation of a temporary magnet. In Figure 3(a), the direction of the flow of electricity is represented by blue dots moving

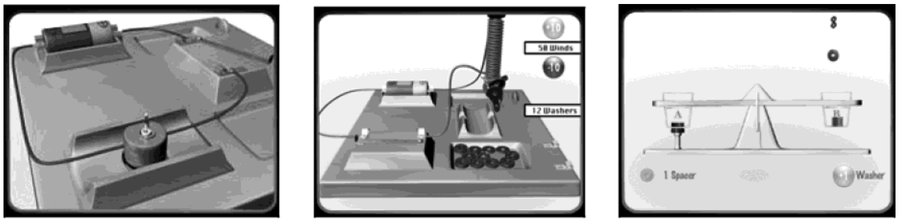


Fig. 4. Examples of interactive animations.

through the wires and bulb and back to the D-cell. The animations enable questions to elicit explanations about what is being shown. As with other concepts and media, the questions become increasingly specific if the students are not expressing an understanding of the point. The animation can also be used to support dialogs in which the student produces an accurate explanation for the events shown; for instance, “You got it! The electricity is flowing through the circuit from the negative to the positive side of the D-Cell.”

Interactive Animations. Interactive Animations allow students to interact directly with the Flash animation through mouse clicks or by using the mouse to move objects on the screen. For example, clicking on the switch in a circuit will open or close the circuit, resulting in a motor running or stopping, or an electromagnet picking up or dropping iron objects (Figure 4). 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. During multimedia dialogs, as students are interacting with a simulation, the tutor can say things like “What could you do to **. . .**? What happens if you ?”

4. DEVELOPING TUTORIAL DIALOGUES

Creating natural and effective interactions between Marni and the student is the overarching goal of the development process. It is necessary to design dialogues 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 dialogues 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 dialogue. 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 we have discussed.

A major challenge of the MyST project was how to design the spoken dialogues and media in a principled way to optimize engagement and learning. To meet this challenge, we developed an iterative approach to dialogue 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. These cycles involved initial human tutoring using a set of illustrations, to human tutoring with computer-based illustrations, animations, and interactive stimulations, to Wizard of Oz studies (described in the following), 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. In addition, we selected a specific approach to tutorial conversations, based on principles of QtA, and then developed, tested and refined dialogs administered first with human tutors, then to initial MyST dialogues monitored and sometime controlled by human tutors in Wizard of Oz sessions. At each step of the development process, sessions were recorded, transcribed and analyzed, leading to refinements and subsequent testing through a series of iterative design and test cycles, and to the final MyST dialogues now being evaluated in schools.

As noted, the concepts addressed in MyST tutorial sessions are aligned with the structure of FOSS content. Each FOSS Module is composed of four investigations, and each investigation consists of a series of four parts. Each of our tutorials is designed to address the key concepts encountered in the individual classroom science investigations for a part of a FOSS investigation. So a FOSS module would have a series of 16 tutorial sessions associated with it (4 investigations of 4 parts each).

4.1. Tutorial Strategy

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. The tutorial dialog is designed to get students to articulate concepts and be able to explain processes underlying their thinking. Tutor actions are designed to encourage students to share what they know and help them articulate why they know what they know. For the system (Marni), the goal of a tutorial session is 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 student. Questioning the Author (QtA) influences the strategies we use to get students to share what they know. QtA is very effective at getting students to think more deeply about a concept. Two of the strategies that it utilizes that are employed by MyST are *marking* and *revoicing*. These two techniques require the ability to identify the student's dialogue content (referred to as marking it) followed by repeating (revoicing) the question back to the student using similar phrasing; for instance, You mentioned that electricity flows in a closed path. What else can you tell me about how electricity flows?

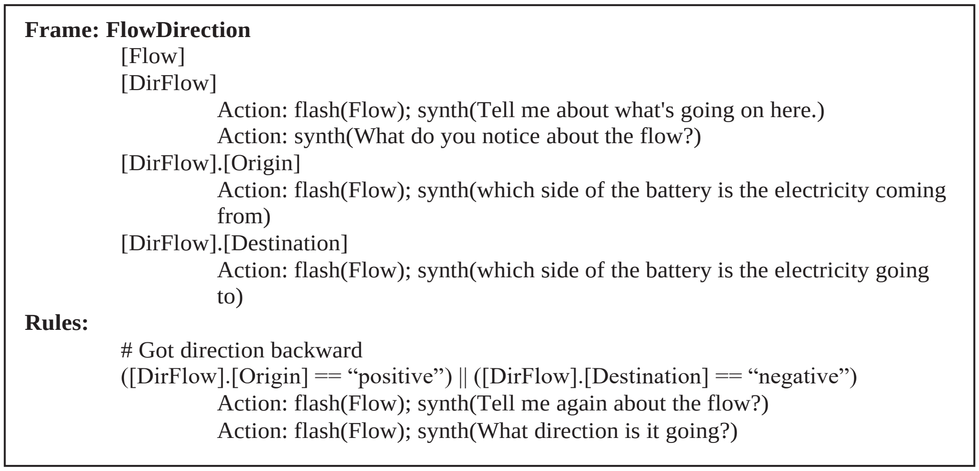
The interactions for a concept typically begin with open-ended questions about the concept. Further sequences are written in such a way that they proceed from more general open-ended questions (What's this all about?) to more directed open-ended questions (Tell me more about the flow of electricity in the circuit). Initially, students are prompted to consider a concept in terms of their recent experiences in class.

4.2. Implementing Tutorial Sessions

Marni's behavior in a dialogue with a student, including the presentation of media within dialogues, is controlled by a *task* file. The *task* file contains the definition of the task frames to be used by the application. A task frame is a data object that contains all of the information necessary (or at least available) to interact about the frame.

—Frame Elements; the extracted information;
—Templates for generating responses;
—Pattern-Action pairs, called Rules, for generating responses contingent on certain conditions in the context.

By default, Marni will attempt to elicit speech to fill the Frame Elements representing the propositions of a frame. A sequence of interface actions is generated to elicit a response. The set of interface actions used are: flash(), movie(), show(), clear(), speak() and synth(). An example action sequence would be *flash(Components); synth(Tell me about that.)*. This sequence would run the Flash file *Components* and would synthesize the word sequence and have Marni speak it. In order to elicit speech to fill a frame element, the system developer specifies a list of action sequences for the element. During a session, the Dialog Manager (DM) keeps count of how many times each element has



```
Frame: FlowDirection
        [Flow]
        [DirFlow]
                Action: flash(Flow); synth(Tell me about what's going on here.)
                Action: synth(What do you notice about the flow?)
        [DirFlow].[Origin]
                Action: flash(Flow); synth(which side of the battery is the electricity coming
                from)
        [DirFlow].[Destination]
                Action: flash(Flow); synth(which side of the battery is the electricity going
                to)
Rules:
        # Got direction backward
        ([DirFlow].[Origin] == “positive”) || ([DirFlow].[Destination] == “negative”)
                Action: flash(Flow); synth(Tell me again about the flow?)
                Action: flash(Flow); synth(What direction is it going?)
```

Fig. 5. Example task frame.

been prompted for and uses the next action sequence in the list. Once it has exhausted the list, it gives the element the value FAIL, and will move on.

The tutorial developer may also specify a set of Rules for the frame. Rules are pattern-action pairs that can be used to generate action sequences conditioned on features of the context. Rule pattern definitions are Boolean expressions based on element values in the context. If the rule evaluates to true, one of the action sequences following it are sent to the interface manager. Like when prompting for an element, the system keeps count of the number of times a rule has been used and uses the next sequence each time. Figure 5 shows an example frame with a rule. The tutor would initially try to elicit information about flow direction by showing an animated Flash file named *Flow* and having the agent say Tell me about what’s going on here. If the student responded with it goes from plus to minus where the direction of electrical flow reversed, the parse would be

[Flow]: [DirFlow].[Origin]: positive [DirFlow].[Destination]: negative

The mapping of *plus* and *minus* to the canonical forms *positive* and *negative* is done by the parser. When the parse is integrated into context, the rule would fire and the tutor would continue to show the flash animation *Flow*, and the avatar would say “*Tell me again about the flow*.”

Rules are also useful for marking and revoicing what students have said. They are used to mark and encourage students to go forward, question students if they get a relationship incorrect, and reward them when their efforts result in responses that accurately express conceptual understandings.

The DM uses a stack driven algorithm for flow control. It maintains two frame stacks, 1) *current*, the set of currently active frames, and 2) *history*, the set of completed frames. The DM tries to complete the frame on top of the *current* stack. If the frame on top is complete, it is moved to the *history* stack and the new top frame is completed. In attempting to complete a frame, the Rules are checked first. If a rule expression evaluates TRUE and it has not been marked FAIL, the next action sequence for the rule is used. If no sequence was generated by checking the Rules, the DM determines the first unfilled frame element that has an associated action sequence. If all required elements are filled, the frame is moved to the *history* stack, and the system attempts to fill the new top frame. The action sequences for both Rules and Frame Elements can cause new frames to be pushed onto the *current* stack, or old frames to be moved off to the *history* stack.

As noted above, development of dialogues, as represented in the *task* files, proceeds through an iterative design, test, and revision process. As new data are received from student sessions, they are analyzed for features like: aspects of the flow of the tutoring session; details of the prompt generation; the use and utility of visuals; and the general completion of frames. This information is used to modify task files to streamline prompts, refine rules, and further design graphics and interactive animations to support or clarify concepts and eliminate misconceptions.

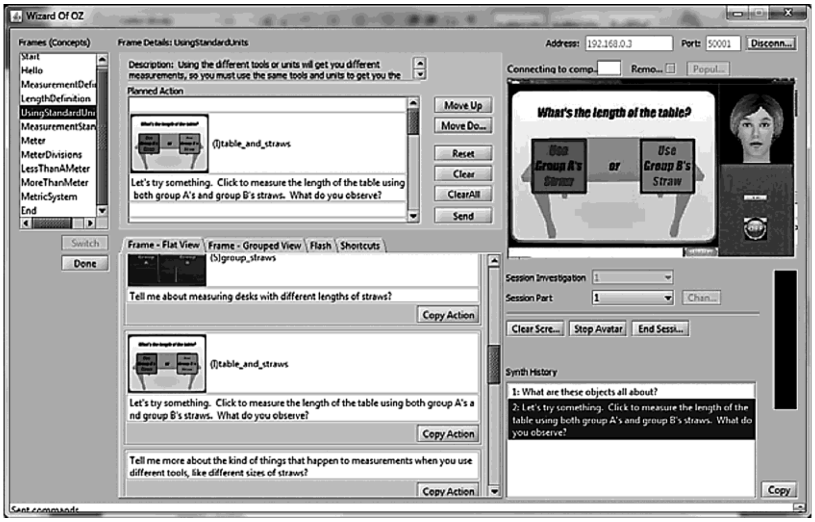
5. WIZARD-OF-OZ INTERFACE

Our development strategy is to model spoken dialogs from 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. For all WOZ data collected, sessions were monitored by project tutors (former science teachers) who served as the Wizards. The data from WOZ sessions was used to improve system coverage 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 have independently interacted with MyST in their schools, while Wizards (either at some other location at the school or at Boulder Language Technologies offices) have monitored the tutoring sessions remotely. One project tutor goes to the school to set up the computers, retrieve students from classrooms, bring them to a computer and initiate the session. The Wizard then connects to a student's MyST session via the internet.

The WOZ interface is a pluggable MyST component. 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. The WOZ system supports both independent use by a student and the ability of a human wizard to connect to any given session. Over the course of the data collection, we have 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%. Fewer than 1% of the wizard interventions involve changing the focus frame. The correct concept was being discussed, but the wizard wanted to say something different.

Wizard display. 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, but can only communicate with the student through the MyST WOZ interface. Figure 6 shows the layout of the Wizard display, which contains the following.

- A screenshot of the screen that the student sees
- The action Marni is about to take
- The frame in focus, including all action sequences associated with elements of the frame



Wizard Of OZ

Frame Details: UsingStandardUnits

Description: Using the different tools or units will get you different measurements, so you must use the same tools and units to get you the

Planned Action

(1)table_and_straws

Let's try something. Click to measure the length of the table using both group A's and group B's straws. What do you observe?

Address: 192.168.0.3 Port: 50001

Session Investigation 1

Session Part 1

(5)group_straws

Tell me about measuring desks with different lengths of straws?

(1)table_and_straws

Let's try something. Click to measure the length of the table using both group A's and group B's straws. What do you observe?

Tell me more about the kind of things that happen to measurements when you use different tools, like different sizes of straws?

Synth History

1: What are these objects all about?

2: Let's try something. Click to measure the length of the table using both group A's and group B's straws. What do you observe?

Fig. 6. Wizard screen.

- A list of all frames in the task file 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. Analysis of log files from WOZ sessions gives insight into problems with tutorials and can lead to development of additional multi-media resources or modification of the task file to cause the system to behave more like the wizards.

Student Interface. An example of the student's screen is shown in Figure 1. The student's computer shows a full screen window that contains the virtual tutor Marni, a display area for presenting information 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. Some displays 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. Push-and-Hold systems work well with children in environments with background noise. Having the hard indication that the user is talking to the system, as compared to an “open mike,” provides useful constraints for the recognizer. 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. It is likely that performance of the speech recognizer was also improved by having the interval of speech indicated by the student.

Dialogue Interaction. 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 dialogue 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 dialogue is oriented around a set of key concepts that the student is expected to know based on the content, instructional activities and learning objectives of each classroom science investigation in each FOSS module. 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 of the dialogue 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 dialogue, the system analyzes utterances produced by a student and maintains a context that represents which points have been addressed by the student, and which have not. In analyzing a student’s answer, the dialog system tests whether the correct entities are filling the correct semantic roles. The dialog manager then generates questions about the missing or erroneous elements to attempt to elicit information about them. The tutor will continue to try to elicit student explanations about an element until the element is filled or the associated prompts are exhausted.

6. MYST SYSTEM ARCHITECTURE

MyST was developed using Boulder Language Technologies Virtual Human Toolkit (VHT). The BLT VHT is a resource for designing and experimenting with multimedia programs that support real time conversational interaction with virtual humans. The VHT provides a general purpose platform, a set of technology modules, and tools for researching and developing conversational systems using natural mixed initiative interaction with users in specific task domains. In mixed-initiative dialogs, either the user or the system can seize the initiative and take control the dialog. The toolkit consists of

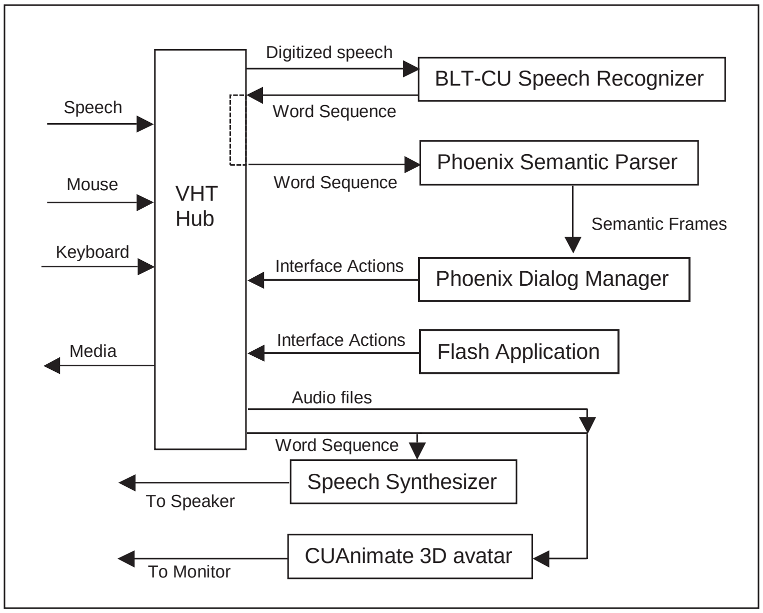


Fig. 7. Virtual human toolkit architecture.

an integrated set of authoring tools and technologies for developing applications that incorporate virtual humans in applications. It provides authoring tools for presenting and interacting with media (text, images, audio, video and animations), designing and controlling lifelike 3D computer characters, and designing natural spoken dialogs with the virtual agent.

The VHT is composed of modules for

—speech recognition;
—speech synthesis;
—semantic parsing;
—dialog management;
—character animation.

It also contains a Hub written in Java that implements the application. The organization of the toolkit is illustrated in Figure 7.

6.1. VHT Hub

The Hub is a Java program that provides all of the functions necessary to invoke and send data to all of the modules, manage the user's input, invoke Flash applications, play media files and invoke the agent. The Hub timestamps and logs all interactions. The Hub executes a set of interaction actions requested by a client module consisting of the following.

—flash(file): execute the specified Flash file;
—movie(file): play the specified media file;
—show(file): display the specified static file;
—clear(): clear the display;
—speak(file): send the prerecorded file to CUAnimate for the character to speak;
—synth(word string): send the specified word string to the TTS then to CUAnimate for the character to speak.

Any client module that implements the Hub Application Program Interface (API) can send interaction requests to the Hub. In Figure 7, both the Phoenix Dialog Manager and a Flash Application are shown sending interaction requests to the Hub. The Dialog Manager can invoke a Flash application, which can in turn use the Hub services.

6.2. Speech Recognizer

The speech recognizer used in the VHT is a large vocabulary continuous speech recognition (LVCSR) system written by Daniel Bolaños [Bolaños et al. 2011], supported jointly by BLT and CU. It uses the general approach of many state-of-the art speech recognition systems: a Viterbi Beam Search is used to find the optimal mapping of the speech input onto a sequence of words. The score for a word sequence is calculated by interpolating Language Model scores and Acoustic Model scores. The Language Model assigns probabilities to sequences of words using trigrams, where the probability of the next word is conditioned on the two previous words. The Language Models were trained using the CMU-Cambridge LM Toolkit [Clarkson and Rosenfeld 1997].

Feature extraction from the audio was carried out using Mel Frequency Cepstral Coefficients (MFCC) plus the logarithm of the signal energy. Cepstral coefficients were extracted using a 20ms window size and a 10ms shift, which produces a feature vector each 10ms that is composed of 12 MFCCs plus the log energy and the first and second order derivatives. Acoustic Models are clustered triphones based on Hidden Markov Models using Gaussian Mixtures to estimate the probabilities of the acoustic observation vectors. The system uses filler models to match the types of disfluencies found in applications. The recognizer can output word graphs, but MyST currently uses only the single best scoring hypothesis. The recognizer is configured to run in approximately real time so the delay after the student quits speaking and before the system is ready to respond is kept short. This is necessary to promote a fluent and engaging dialog.

6.3. Semantic Parser

The Phoenix parser [Ward 1994] maps the speech recognizer output, or any text, onto a sequence of semantic frames. These frames represent the system's understanding of an utterance. The type of representation Phoenix uses to extract information from user input is generally referred to as shallow semantics. Shallow semantics represents the entities, events and relations between them important to understanding an utterance. In Phoenix, these are characterized as semantic frames, together with semantic frame elements. An example parse for *Electricity goes from minus to plus* is:

Frame: FlowDirection
[**Electricity**] (electricity)
[**Flows**] (goes)
[**DirFlow**].[**Origin**] Negative(minus)
[**DirFlow**].[**Dest**] Positive(plus)

Semantic grammars are used to match word strings against patterns for frame elements. These are Context Free patterns where the NonTerminals are concepts, events and relations important in the domain. Separate grammars are written for each Frame Element (like [DirFlow].[Origin]). In matching Frame Element grammar patterns against the input text, the parser ignores words that do not match any frame element. This allows the system to match expressions relevant to understanding the domain while ignoring extraneous information and disfluencies such as restarts. A Viterbi search is used to find the optimal set of frames and frame elements. The most optimal parse is the one that covers most of the input and is least fragmented. A set of parses of equal score is produced for an ambiguous input. The grammar rules may be written manually or may be trained from an annotated corpus if one is available.

6.4. Dialog Manager

The Dialog Manager controls the system's dialogue interaction with the user and is responsible for:

(a) maintaining a context representing the history of the dialog;
(b) selecting a preferred parse from a set of candidate parses given the context;
(c) integrating the new parsed input into the context;
(d) generating a sequence of actions based on the context.

The DM also uses the frame representation used by the parser. It also provides a mechanism for developers to specify the behavior of the system. This mechanism was discussed in Section 4.

6.5. Character Animation

Within the toolkit, a set of ethnically diverse animated agents each produce anatomically correct visual speech (through movements of the lips, tongue, and jaw) synchronized automatically with either recorded speech (given a text string representing the spoken words) or with synthesized speech generated by a text-to-speech synthesis program. The CU Animate [Ma et al. 2002, 2004] module enables authors to produce facial expressions and animation sequences during speech production, while "listening" to the user, or in response to mouse clicks or other input modes. Each animated agent can produce accurate facial expressions of six basic emotions (surprise, joy, sadness, fear, disgust, anger). In MyST, the character for Marni shown in Figure 1 was used in all applications.

6.6. Text-to-Speech Synthesis

A Text-To-Speech synthesizer receives word strings from the natural language generator and synthesizes them into audio waveforms that can be played back to the user. The VHT interfaces to the general-purpose Festival speech synthesis system [Taylor et al. 1998], and to the commercially available Acapela synthesizer.

7. USE OF SPOKEN RESPONSES

In the tradition of other systems using children's speech [Mostow and Aist 1999], 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. When these occur, the system asks a follow-on question, which may be accompanied by a new illustration, animation or interactive investigation, that is designed to scaffold learning and elicit an appropriate response. Thus, the interaction style used in Questioning the Author is especially well suited to ASR errors that can occur during spoken dialogues.

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 dialogue 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 dialogue, the system will conclude the session with a default tutorial presentation as specified in the *task* file for the session.

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 dialogue 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 can catch some 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.

8. 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 corpus can be used to train and evaluate children's speech recognition and spoken dialog algorithms. It can also be used to support analyses of the characteristics of children's speech. We are also using the data for modeling tutorial dialogs 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 are being 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.

Data are being collected in three basic conditions.

(1) *Human Tutor.* A human tutor conducts a tutorial with a student. The human tutor has access to the visuals and other supplementary materials, but the tutor talks directly with the student.
(2) *Wizard-Of-Oz.* The WOZ interface is used to interact with the student as described earlier. All interactions are saved to a time-stamped log file.
(3) *Stand-alone Virtual Tutor.* Students interact with the MyST system without a wizard being connected. This is the procedure being used to assess the effectiveness of the MyST system in schools. Data collection in this condition recently started and is not included in Table I, or in any of the data sets used in this paper.

Table I shows the amount of data (number of speakers and hours of speech) collected for each module. The Water module was developed last and collection is just beginning.

Speech Files. The speech data are stored in files by student turns, that is, whatever is said from the time the student pressed the space bar to talk until the bar is released.

Table I. Data Collected by Module

Module	All		Training		Development		Evaluation	
	speakers	hours	speakers	hours	speakers	hours	speakers	hours
Magnetism and Electricity	176	35	149	31	14	2	13	2
Measurement	222	48	185	38	20	5	17	5
Variables	60	20	44	18	6	1	10	1
Water	25	8	22	6	1	1	2	1
Total	483	111	400	93	41	9	42	9

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. Thus far, 111 hours of speech have been transcribed.

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 offline 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.

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.

9. COMPONENT EVALUATIONS

Since only a small amount of data has been collected for the Water (WA) module, and transcripts for those are not completed, experiments were conducted using data from only 3 modules; Magnetism & Electricity (ME), Measurement (MS), and Variables (VB). As shown in Table I, the data were partitioned by speaker into training, development and evaluation sets. Data from any individual 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.

9.1. Automatic Speech Recognition Performance

The recognizer was 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 6235 words. The Word Error Rate (WER) for the recognizer on the Evaluation set is shown in Table II 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.

Table II. 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

The WER for the pooled data (Tot) was 30.9%. For the individual modules, the WER for ME and MS were very similar, while the WER for VB was substantially higher. Using a global LM, the perplexity of each module was: 56 for ME, 63 for MS and 74 for VB. Even though the ME data had a lower perplexity than the MS, the WERs are similar. VB had a substantially higher perplexity and WER. The higher perplexity of the VB data can be attributed both to less training data and to the topic of the module. The ME and MS modules are about concrete topics with which students are generally familiar. Variables introduces more abstract ideas like dependent and independent variables and graphing data. Students generally have a more difficult time with this topic, even with human tutors.

The baseline results reported above 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 and Woodland 1995] and Vocal Track Length Normalization [Lee and 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 format 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 covariances 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. A regression class tree is used to cluster the Gaussian components in the system. The number of base classes in the tree was set to 50 and the tree was built using EM clustering. Full transformation matrices for the means and diagonal transformation matrices for the variances were used. The minimum class occupancy count was set to 3500. Adding MLLR adaptation reduced the error rate further to 27.4%.

For the numbers we have listed, 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 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 precompute warp factors

Table III. Speech Recognition Performance by LM

	Overall			ME			MS			VB		
	WER (%)	PP	CA	WER (%)	PP	CA	WER (%)	PP	CA	WER (%)	PP	CA
Global	27.4	63	.86 .90	26.1	56	.87 .91	26.7	63	.86 .89	31.9	74	.82 .90
ModA	27.8	62	.87 .90	26.3	55	.88 .91	27.1	63	.87 .89	32.0	73	.83 .89
ModS	27.9	60	.88 .89	26.3	53	.89 .90	27.1	59	.87 .88	32.0	73	.86 .87
InvA	27.2	61	.87 .89	26.0	54	.89 .91	26.3	62	.87 .89	31.7	72	.85 .88
InvS	29.2	64	.89 .87	27.8	56	.90 .89	28.3	64	.88 .87	34.0	75	.86 .85

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 II, 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.

Effect of Language Model Specificity. The VHT speech decoder uses standard trigram language models that were trained using the CMU-Cambridge Language Model Toolkit [Clarkson and Rosenfeld 1997]. In creating language models for structured data such as this, the developer has the opportunity to tune the model to the specific topic of the investigations. In this case, a general language model is trained and adaptation data is used to tune the model for a specific topic. One effective method for language model adaptation is to use MAP (maximum a posteriori) adaptation, which combines weighted word counts from the general data and adaptation data [Federico 1996]. We used a simple approximation to this procedure by mixing adaptation data with general data with a weighting factor. The weighting factor was determined using a development set. Performance of the recognizer was determined using five sets of Language Models (LMs).

(1) *Global*. A single LM is trained by pooling all training data;
(2) *ModA*. A separate LM is generated for each module by adapting the Global model with the training data for the module;
(3) *ModS*. A separate LM is trained for each module using just the training data for the module;
(4) *InvA*. A separate LM is generated for each investigation by adapting the Global model with the training data for the investigation;
(5) *InvS*. A separate LM is trained for each investigation using just the training data for the investigation.

The WER in %, Perplexity (PP) and Concept Accuracy (CA) for modules in the Evaluation set are shown in Table III. For CA, the top number is Recall and the bottom number is Precision. The WER, PP and CA numbers for investigation specific models are an average across the investigations of each module. There is a small difference in WER as the LMs become more specific. Results with the Module Adapted (ModA) and Module Specific (ModS) LMs are substantially equivalent and are slightly worse than the WER achieved with the Global LM. The Investigation Adapted (InvA) LMs had, for each module and overall, lower WER than the Investigation Specific (InvS) LMs which had the highest WER. The data were not clearly sufficient to train investigation specific LMs, but LM adaptation helped a little bit in this case, although not enough to ensure a significant improvement with respect to the WER achieved with the Global LM. Variations in perplexity across LMs are also small.

Disfluencies. Conversational speech contains many events that are nonwords, such as breath and filled pauses. A common technique to deal with these events is to use acoustic filler models to match the input. In addition to a Silence model, the system uses acoustic models to match nonword speech events (br, EM, HMM, HUH, MMM, UHM). The decoder that produced the results in Tables II and III used filler models. Fillers are allowed to occur between any events (words or other fillers) with an insertion penalty that is set to minimize WER using the development set. We conducted an investigation to give some information about the performance of the filler models used in the system. Using a global language model, the overall WERs of the baseline system and the adapted system were 30.9% and 27.4%, respectively. Approximately 6.7% of the annotated tokens in the evaluation set transcriptions were fillers. Filler tokens are stripped out of the recognition hypotheses before computing WER and before parsing, so insertions of filler tokens do not in themselves cause a problem. A problem can occur when recognizing the filler causes a word deletion or substitution error. Without using filler models the WERs increased to 35.1% and 29.3%. It was clearly beneficial to overall WER to include filler models in the decoder. Even using filler models, disfluencies continue to be a significant problem in ASR for children's conversational speech.

9.2. Concept Accuracy

The behavior of the virtual tutor is more dependent on Concept Accuracy than on Word Error Rate. The only representation that the Dialog Manager has of what the student said are the extracted frames produced by the parser. If two different word strings have the same parser output, then they are equivalent to the Dialog Manager. 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 (there are 4 investigations for each module) 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 Tables II and III. The first (or top) number in the accuracy is Recall and the second (or bottom) number is Precision. As seen in Table II, 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. This generally would be the expected effect of recognizing more content words correctly. As seen in Table III, increasing the specificity of the LM results in an increase in Recall at the expense of a decrease in Precision. This trend can be explained by realizing that more specific LMs tend to increase the likelihood that domain specific content words will be recognized, whether they were actually spoken or not. This expectation is consistent with the CA results.

10. STUDENTS' AND TEACHERS' IMPRESSIONS OF MYST

A written survey was given to 167 students who used MyST in five elementary schools during the 2009–2010 school year. All of these students used MyST in WOZ mode. 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

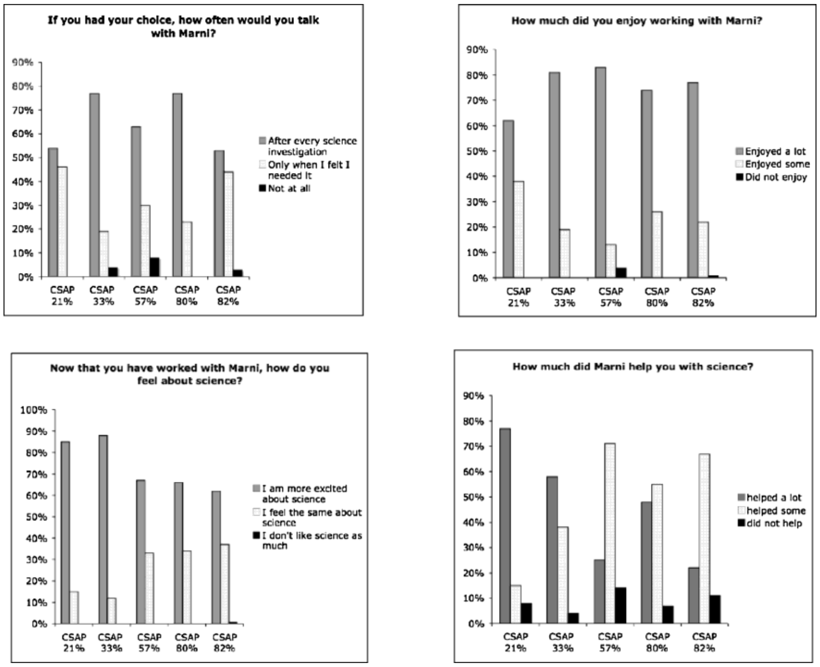


Fig. 8. Student survey responses by school CSAP score.

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 ten 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. Four main questions assessed student experiences with Marni: 1) How much did Marni help you with science? 2) How much did you enjoy working with Marni? 3) If you had your choice, when would you talk with Marni? 4) Now that you have worked with Marni, how do you feel about science? In addition, several other questions were included to assess usability issues, such as "Did you understand Marni's voice?"

The schools in which students used MyST varied greatly in terms of the percentage of students who scored proficient or above in science on the state Colorado Student Assessment Program (CSAP) test: from 21% proficient or above for the lowest scoring school, to 82% proficient or above in the highest scoring school. Figure 8 displays the distribution of students' response choices to each question. The histograms are grouped by school, using the percentage of students at the school who scored as proficient or above. In general, students had positive experiences and impressions about the program. Across schools, 50% to 75% of students said they would like to talk with Marni after every science investigation, 60% to 80% said they enjoyed working with Marni "a lot," and 60% to 90% selected "I am more excited about science" after using the program. Perhaps most interesting, the majority of students in the lowest two performing schools felt that Marni "helped a lot" in learning science (75%, 55%), whereas the majority of students in the higher performing schools responded that Marni "helped some." Since MyST dialogs are designed to help students learn the science concepts embedded in

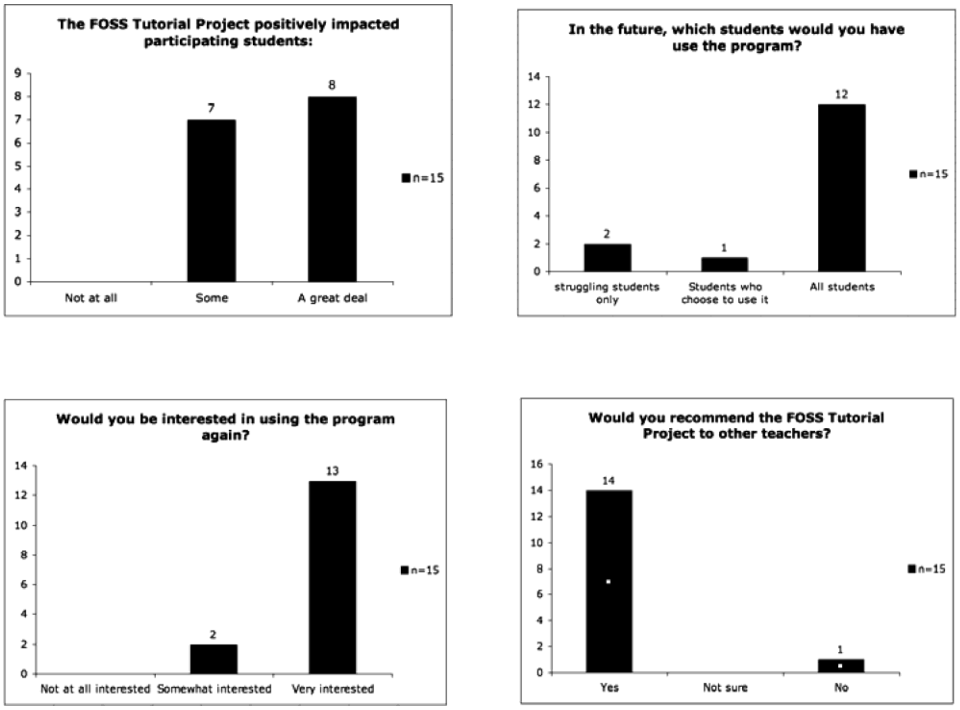


Fig. 9. Teacher survey results.

classroom investigations, MyST should provide the most benefit to students who are having difficulty understanding these concepts. The survey responses produce initial evidence that students who have most to gain from using MyST have more positive impressions of the program.

Teachers were asked for feedback to help assess the feasibility of an intervention using the system and their perceptions of the impact of the system. 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 results of the survey are shown in Figure 9. Even though students who used MyST left the classroom during tutoring sessions, teacher responses were in general very positive. 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 benefited students who did not use the system. Teachers indicated that they would like to have all of their students use the system (not just struggling students) and that they would recommend it to other teachers.

11. CONCLUSIONS AND FUTURE WORK

This article has presented the design of a conversational multimedia virtual tutor for elementary school science. Speech and language technologies play a central role because the focus of the system is on engagement and self-expression by the students. It was argued that current speech and natural language technology is a good match to this task because it takes advantage of speech understanding capabilities to improve the interaction while minimizing the effects of errors in recognition and understanding.

A corpus is being developed which will be used to evaluate the MyST system as well as enable research by others on tutorial dialog systems. Evaluation results were presented for the Automatic Speech Recognition and Spoken Language Understanding components of the system. Using a global LM, the baseline system had a WER of 30.9% with an overall Recall of 84% and Precision of 89% for extraction of frame elements. With batch unsupervised speaker adaptation, a WER of 27.4% with a Recall of 86% and a Precision of 90% were achieved. The accuracy of extraction of frame elements measures how well the system is understanding the student. Performance of live systems would average somewhere between these performance numbers.

During data collection using a WOZ paradigm, surveys were collected from students and teachers that bear on the engagement and feasibility of the proposed tutoring system. Following a series of tutoring sessions with Marni, the great majority of students reported that they enjoyed spending time working with her, that they felt that Marni helped them learn science, and perhaps most interesting, that they felt more interested in science and more motivated to learn science than they had before using the system. Students in both high performing and low performing schools, the latter including significant populations of English language learners and students from families with low socioeconomic status, reported that Marni was “way cool.” One of the unanticipated benefits of this shared perception to our project was that students whose parents did not 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.

The third, fourth, and fifth grade teachers whose students were tutored by Marni were also excited about the program. The teachers noticed that most of their students who used the program increased their participation and contributions during science investigations and classroom discussions, and this benefited all students, including those who were not being tutored. Teachers reported that they would like to use MyST in the future to tutor all of their students, and that they would recommend the program to other teachers.

The survey responses reported in this article are based on experiences with a WOZ system. Students interacted with a virtual tutor, but a human tutor was moderating the interaction. Survey responses and anecdotal evidence in observing interactions indicate that both students and teachers are accepting of the virtual tutor. What remains to be shown is how well the virtual tutor is able to maintain engagement without the assistance of a wizard. The efficacy of the system in the form of learning gains also needs to be determined. At the time of this writing, during the 2010–2011 school year, MyST is being evaluated in stand-alone mode. In addition to student and teacher surveys, the system is being evaluated for its potential to improve student achievement during independent use by children in each of the four areas of science. In the evaluation phase of the project, children in classrooms (whose parents consent to their child being tutored) are randomly assigned to one of two groups: being tutored by Marni, or being tutored in small groups by one of the project tutors trained in QtA who tutored children and served as Wizards in the development phase of the project. ASK assessments are given to students before and after each science module. Gains in science learning will be compared for students in these two groups based on their performance on the ASK assessment administered to each student before and after each science module. In addition, the performance of these students will be compared to the performance of students who are administered ASK assessments in classrooms that did not receive tutoring. Our hypothesis is that students who engage in multimedia dialogs with Marni will produce benefits similar to students who interact with human tutors. One of the most important outcomes of the assessment procedure will be determining the feasibility and potential of using speech and language processing technologies in multimedia tutoring dialogs with children.

REFERENCES

Aist, G. and Mostow, J. 2009. Designing spoken tutorial dialogue with children to elicit predictable but educationally valuable responses. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech).*

Atkinson, R. K. 2002. Optimizing learning from examples using animated pedagogical agents. *J. Educ. Psych.* 94, 416–427.

Baylor, A. L. and Ryu, J. 2003. Does the presence of image and animation enhance pedagogical agent persona? *J. Edu. Comput. Resear., 28*, 4, 373–395.

Baylor, A. L. and Kim, Y. 2005. Simulating instructional roles through pedagogical agents. *Int. J. Artific. Intell. Edu. 15*, 1.

Beck, I. L., McKeown, M. G., Worthy, J., Sandora, C. A., and Kucan, L. 1996. Questioning the author: A year-long classroom implementation to engage students with text. *Elem. School J. 96*, 4, 387–416.

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

Bernstein, J. and Cheng, J. 2007. Logic and validation of fully automatic spoken English test. In *The Path of Speech Technologies in Computer Assisted Language Learning: From Research Toward Prac- tice,* M. Holland and F. P. Fisher, Eds., Routledge, 174–194. http://www.ordinate.com/samples/Versant-English/Sample-TEST-PAPER-Versant-English-Test-watermark.pdf

Bloom, B. S. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. *Educ. Resear. 13*, 4–16.

Bolanos, D., Cole, R., Ward, W., Borts, E., and Svirsky, E. 2011. FLORA: Fluent oral reading assessment of children's speech. *ACM Trans. Speech Lang. Process.*

Bruner, J. S. 1966. *Toward a Theory of Instruction*. Harvard University Press, Cambridge, MA.

Bruner, J. S. 1990. *Acts of Meaning*. Harvard University Press, Cambridge, MA.

Butcher, K. R. 2006. Learning from text with diagrams: Promoting mental model development and inference generation. *J. Edu. Psych. 98*, 1, 182–197.

Chapin, S. H., O'Connor, C., and Anderson, N. C. 2003. *Classroom Discussions Using Math Talk to Help Students Learn*. Math Solution Publications, Sausalito, CA.

Chen, W., Mostow, J., and Aist, G. 2010. Exploiting predictable response training to improve automatic recognition of children's spoken questions. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems (ITS2010)*, Springer-Verlag, 55–64.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., and Glaser, R. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cogn. Sci. 13*, 145–182.

Chi, M. T. H., De Leeuw, N., Chiu, M., and Lavancher, C. 1994. Eliciting self-explanations improves understanding. *Cogn. Sci. 18*, 439–477.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., and Hausmann, R. G. 2001. Learning from human tutoring. *Cogn. Sci. 25,* 471–533.

Clarkson, P. R. and Rosenfeld, R. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of Eurospeech*.

Cohen, P. A., Kulik, J. A., and Kulik, C. L. C. 1982. Educational outcomes of tutoring: A meta-analysis of findings. *Am. Edu. Resear. J. 19*, 237–248.

Cole, R., Van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W., and Yan, J. 2003. Perceptive animated interfaces: First steps toward a new paradigm for human–computer interaction. In*Proc. IEEE 91*, 9, 1391–1405.

Cole, R., Wise, B., and Van Vuuren, S. 2007. How Marni teaches children to read. *Educ. Techn.*

Craig, S. D., Gholson, B., Ventura, M., Graesser, A. S., and Tutoring Research Group. 2000. Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *Int. J. Artif. Intell. Edu. 11,* 242–253.

Dede, C., Salzman, M., Loftin, B., and Ash, K. (in press). *Using virtual reality technology to convey abstract scientific concepts*. In *Learning the Sciences of the 21st Century: Research, Design, and Implementing Advanced Technology Learning Environments*. M. J. Jacobson and R. B. Kozma, Eds. Lawrence Erlbaum, Hillsdale, NJ.

Driscoll, D., Craig, S. D., Gholson, B., Ventura, M., Hu, X., and Graesser, A. 2003. Vicarious learning: Effects of overhearing dialog and monolog-like discourse in a virtual tutoring session. *J. Educ. Comput. Resear. 29,* 431–450.

Federico, M. 1996. Bayesian Estimation Methods for n-gram language model adaptation. In *Proceedings of ICSLP'96*, 240–243.

Graesser, A. C., Hu, X., Susarla, S., Harter, D., Person, N. K., Louwerse, M., Olde, B. and the Tutoring Research Group. 2001. AutoTutor: An intelligent tutor and conversational tutoring scaffold. In *Proceedings of the 10th International Conference of Artificial Intelligence in Education,* 47–49.

Graesser, A., N., Person, N., and Harter D. 2001. Teaching tactics and dialog in Autotutor. *Int. J. Artific. Intell. Edu.*

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

Hausmann, R. G. M. and Vanlehn, K. 2007b. Self-explaining in the classroom: Learning curve evidence. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. D. McNamara and G. Trafton Eds., Erlbaum, Mahwah, NJ, 1067–1072.

King, A. 1989. Effects of self-questioning training on college students' comprehension of lectures. *Contemp. Educ. Psy. 14*, 366–381.

King, A. 1991. Effects of training in strategic questioning on children's problem-solving performance. *J. Educ. Psych. 83*, 307–317.

King, A. 1994. Guiding knowledge construction in the classroom: Effect of teaching children how to question and explain. *Am. Educ. Resear. J. 31,* 338–368.

King, A., Staffieri, A., and Adelgais, A. 1998. Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *J. Educ. Psych. 90,* 134–15.

Kintsch, W. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psych. Rev. 95,* 163–182.

Kintsch, W. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, England.

Lee, L. and Rose, R. C. 1998. A frequency warping approach to speaker normalization. *IEEE Trans. Speech Audio Process. 6*, 1, 49–60.

Leggetter, C. J. And Woodland, P. C. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models. *Comput. Speech Langu. 9*, 171–185.

Lester, J., Converse, S., Kahler, S., Barlow, S., Stone, B., and Boghal, R. 1997. The persona effect: Affective impact of animated pedagogical agents. In *Proceedings of CHI'97, ACM,* New York, 359–366.

Lester, J., Stone, B., and Stelling. G. 1999. Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments. *User Model. User-Adap. Interact. 9*, 1–2, 1–44.

Littman, D. and Silliman, X. 2004. ITSPOKE: An intelligent tutoring spoken dialog system. In *Proceedings of HLT-NAACL*, 5–8.

Ma, J., Cole, R. A., Pellom, B., Ward, W., and Wise, B. 2004. Accurate automatic visible speech synthesis of arbitrary 3d models based on concatenation of di-viseme motion capture data. *J. Comput. Anim. Virt. Worlds 15*, 485–500.

Ma, J. Yan, J., and Cole, R. 2002. CU Animate: Tools for enabling conversations with animated characters. In *Proceedings of the International Conference on Spoken Language Processing*.

Madden, N. A. and Slavin, R. E. 1989. Effective pullout programs for students at risk. in *Effective Pro- grams for Students At Risk*, R. E. Slavin, N. L. Karweit, and N. A. Madden, Eds., Allyn and Bacon, Boston, MA.

Mayer, R. 2001. *Multimedia Learning*. Cambridge University Press, Cambridge, UK.

McKeown, M. G. and Beck, I. L. 1999. Getting the discussion started. *Educ. Leader. 57*, 3, 25–28.

McKeown, M. G., Beck, I. L., Hamilton, R., and Kucan, L. 1999. *Accessibles—Questioning the Author (Easy-Access Resources for Classroom Challenges).* Wright Group, Bothell, WA.

Moreno, R., Mayer, R. E., Spires, H. A., and Lester, J. C. 2001. The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cogn. Inst. 19*, 2, 177–213.

Mostow, J. and 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 J. 16*, 3, 407–424.

Mostow, J. and Aist, G. 2001. Evaluating tutors that listen: An overview of Project LISTEN. In *Smart Machines in Education*, K. Forbus and P. Feltovich, Eds.

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Huang, C., Junker, B., Sklar, M. B., and Tobin, B. 2003. Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *J. Educa. Comput. Resear. 29*, 1, 61–117.

Mostow, J. and Chen, W. 2009. Generating instruction automatically for the reading strategy of self-questioning. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED'09).* 465–472.

Murphy, P. K. and Edwards. M. N. 2005. What the studies tell us: A meta-analysis of discussion approaches. In *Making Sense of Group Discussions Designed to Promote High-Level Comprehension of Texts. Symposium Presented at the Annual Meeting of the American Educational Research Association*.

Murphy, P. K., Wilkinson, I. A. G., Soter, A. O., Hennessey, M. N., and Alexander, J. F. 2009. Examining the effects of classroom discussion on students' high-level comprehension of text: A meta-analysis. *J. Educ. Psych. 101*, 740–764.

Naep. 2002. http://nces.ed.gov/nationsreportcard

Nass C. and Brave S. 2005. Wired for Speech: How Voice Activates and Advances The Human-Computer Relationship. MIT Press, Cambridge, MA.

Nystrand, M. and Gamoran, A. 1991. Instructional discourse, student engagement, and literature achievement. *Resear. Teach. English 25,* 261–290.

Palincsar, A. S. 1998. Social constructivist perspectives on teaching and learning. *Annual Revi. Psych. 49*, 345–375.

Palincsar, A. S. and Brown, A. 1984. Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cogn. Instr. 1,* 117–175.

Pine, K. J. and Messer, D. J. 2000. The effect of explaining another's actions on children's implicit theories of balance. *Cogn. Instr. 18*, 1, 35–51.

Reeves, B. and Nass, C. 1996. *The Media Equation,* Cambridge University Press, Cambridge, UK.

Rickel, J. and Johnson, W. L. 2000. Task-oriented collaboration with embodied agents in virtual worlds. In *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds.

Soter, A. O. and Rudge, L. 2005. What the discourse tells us: Talk and indicators of high-level comprehension. In *Proceedings of the Annual Meeting of the American Educational Research Association*. 11–15.

Soter, A. O., Wilkinson, I. A. G., Murphy, P. K., Rudge, L., Reninger, K., and Edwards, M. 2008. What the discourse tells us: Talk and indicators of high-level comprehension. *Int. J. Educ. Resear. 47*, 372–391.

Taylor, P., Black, A. W., and Caley, R. 1998. The architecture of the festival speech synthesis. In *Proceedings of the 3rd ESCA Workshop in Speech Synthesis*. 147–151.

Topping, K. and Whitley, M. 1990. Participant evaluation of parent-tutored and peer-tutored projects in reading, In *Educa. Resear. 32*, 1, 14–32.

Van Lehn, K. and Graesser, A. C. 2002. Why2 Report: Evaluation of Why/Atlas, Why/AutoTutor, and accomplished human tutors on learning gains for qualitative physics problems and explanations. Unpublished report prepared by the University of Pittsburgh CIRCLE group and the University of Memphis Tutoring Research Group.

Van Lehn, K., Lynch, C., Taylor, L., Weinstein, A., Shelby, R., Schulze, K., Treacy, D., and Wintersgill, M. 2003. In *Intelligent Tutoring Systems,* S. A. Cerri, G. Gouarderes, and F. Paraguacu, Eds. Springer, Berlin, Germany, 367–376.

Vanlehn, K., Lynch, C., Schulze, K. Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., and Wintersgill, M. 2005. The Andes physics tutoring system: Five years of evaluations. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education*. G. McCalla and C. K. Looi, Eds. IOR Press, Amsterdam.

Vygotsky, L. S. 1978. *Mind in Society: The Development of Higher Psychological Processes.* Harvard University Press, Cambridge, MA.

Ward, W. 1994. Extracting information from spontaneous speech, In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*

Ward, W. and Pellom, B. 1999. The CU Communicator system. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*.

WISE, B., COLE, R., VAN VUUREN, S., SCHWARTZ, S., SNYDER, L., NGAMPATIPATPONG, N., TUANTRANONT, J., AND PELLOM, *B.* 2005. Learning to read with a virtual tutor: foundations to literacy. In *Interactive Literacy Education,* C. Kinzer and L. Verhoeven, Eds., Environments through Technology. Lawrence Erlbaum, Mahwah, NJ.

WOOD, D. AND MIDDLETON, D. 1975. A study of assisted problem solving. *Brit. J. Psych. 66,* 181–191.

Received June 2010; revised October 2010 and April 2011; accepted April 2011

Journal of Educational Psychology

My Science Tutor: A Conversational Multimedia Virtual Tutor

Wayne Ward, Ron Cole, Daniel Bolaños, Cindy Buchenroth-Martin, Edward Svirsky, and Tim Weston

Online First Publication, September 9, 2013. doi: 10.1037/a0031589

CITATION

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., & Weston, T. (2013, September 9). My Science Tutor: A Conversational Multimedia Virtual Tutor. *Journal of Educational Psychology*. Advance online publication. doi: 10.1037/a0031589

Journal of Educational Psychology

© 2013 American Psychological Association
0022-0663/13/$12.00 DOI: 10.1037/a0031589

My Science Tutor: A Conversational Multimedia Virtual Tutor

Wayne Ward
Boulder Language Technologies, Boulder, Colorado, and
University of Colorado at Boulder

Ron Cole, Daniel Bolaños,
Cindy Buchenroth-Martin, and Edward Svirsky
Boulder Language Technologies

Tim Weston
University of Colorado at Boulder

My Science Tutor (MyST) is an intelligent tutoring system designed to improve science learning by elementary school students through conversational dialogs with a virtual science tutor in an interactive multimedia environment. Marni, a lifelike 3-D character, engages individual students in spoken dialogs following classroom investigations using the kit-based Full Option Science System program. MyST attempts to elicit self-expression from students; process their spoken explanations to assess understanding; and scaffold learning by asking open-ended questions accompanied by illustrations, animations, or interactive simulations related to the science concepts being learned. MyST uses automatic speech recognition, natural language processing, and dialog-modeling technologies to interpret student responses and manage the dialog. Sixteen 20-min tutorials were developed for each of 4 areas of science taught in 3rd, 4th, and 5th grades. During summative evaluation of the program, students received one-on-one tutoring via MyST or an expert human tutor following classroom instruction on the science topic, representing over 4.5 hr of tutoring across the 16 sessions. A quasi-experimental design was used to compare average learning gain for 3 groups: human tutoring, virtual tutoring, and no tutoring. Learning gain was measured using standardized assessments given to students in each condition before and after each science module. Results showed that students in both the human and virtual tutoring groups had significant learning gains relative to students in the control classrooms and that there were no significant differences in learning gains between students in the human and MyST human tutoring conditions. Both teachers and students gave high-positive survey ratings to MyST.

Keywords: intelligent tutors, spoken dialog, science learning

According to the 2009 National Assessment of Educational Progress (NAEP, 2005), only 34% of fourth graders, 30% of eighth graders, and 21% of 12 graders tested as proficient in science, with 1%–2% of these students demonstrating advanced knowledge of science in these grades. Thus, over two thirds of U.S. students are not proficient in science. The vast majority of these students are in low-performing schools that include a high percentage of disadvantaged students from families with low socioeconomic status, which often include English learners with low English-language proficiency. Analysis of the NAEP scores in reading, math, and science over the past 20 years indicate that this situation is getting worse. For example, the gap between English learners and English-only students, which is over one standard deviation lower for English learners, has increased rather than decreased over the past 20 years. Moreover, science instruction is often underemphasized in U.S. schools, with reading and math being stressed. My Science Tutor (MyST) was designed to address this problem by immersing students in a multimedia environment with a virtual science tutor that was designed to behave like an engaging and effective human tutor. The focus of the program is to improve each student's engagement, motivation, and learning by helping them learn to visualize, reason about, and explain science during conversations with the virtual tutor.

The learning principles embedded in MyST are consistent with conclusions and recommendations of the National Research Council Report, "Taking Science to School: Learning and Teaching Science in Grades K-8" (Duschl, Schweingruber, & Shouse,

Wayne Ward, Boulder Language Technologies, Boulder, Colorado, and Computational Language and Education Research Center, University of Colorado at Boulder; Ron Cole, Daniel Bolaños, Cindy Buchenroth-Martin, and Edward Svirsky, Boulder Language Technologies; Tim Weston, ATLAS Center, University of Colorado at Boulder.

The research reported here was supported by Institute of Education Sciences, U.S. Department of Education Award R305B070008, National Science Foundation Grants DRL 0733323, awarded to Boulder Language Technologies, and DRL 0733322, awarded to the University of Colorado at Boulder, and NIH Award R44 HD055028 to Mentor Interactive Inc. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute of Education Sciences or the National Science Foundation. We gratefully acknowledge the help of Angel Stobaugh, Director of Literacy Education at Boulder Valley School District, and the principals and teachers who allowed us to visit their schools and classrooms. We appreciate the amazing efforts of Jennifer Borum, Linda Hill, Suzan Heglin, and the rest of the human experts who scored the text passages.

Correspondence concerning this article should be addressed to Wayne Ward, Boulder Language Technologies, 2960 Center Green Court, Boulder, CO 80301. E-mail: wward@bltek.com

2007), which emphasizes the critical importance of scientific discourse in K–12 science education. The report identifies the following 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. (p. 2)

The report also emphasizes 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.

In a meta-analysis of 18 studies by Chi (2009), the author examined student learning along the continuum *active, constructive, interactive*. Active tasks include "doing something," such as participating in a classroom science investigation. Constructive tasks include "producing something," such as a written report describing the results of the investigation. Interactive tasks require discourse and argumentation with a peer or tutor. Chi's analysis of the research studies produced strong evidence that interactive tasks produce the greatest learning gains.

A substantial body of research indicates that engaging in discourse and argumentation about science is one of the most challenging tasks for young learners, and one of the most important and beneficial skills for them to acquire (Hake, 1998; Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009; Osborne, 2010; Soter et al., 2008). However, evidence also indicates that authentic conversations are extremely rare across all content areas in U.S. classrooms (Cazden, 1988; Gamoran & Nystrand, 1991; Nystrand, 1997). As Osborne (2010) noted, "Argument and debate are common in science, yet they are virtually absent in science education" (p. 463). Our goal in designing MyST was to provide students with the scaffolding, modeling, and practice they need to learn to reason and talk about science.

MyST is an intelligent tutoring system intended to provide an intervention for third-, fourth-, and fifth-grade children who are struggling with science. In our study, it was used as a supplement to normal classroom instruction using the Full Option Science System (FOSS). FOSS is an inquiry-based science program that is based on the idea that "The best way for students to appreciate the scientific enterprise, learn important scientific concepts, and develop the ability to think well is to actively construct ideas through their own inquiries, investigations and analyses" (FOSS, n.d., para. 3). It has been under development since 1988, and is in use in every state in the United States. Twenty-six science modules have been developed for Grades K– 6. The learning objectives in each FOSS module are aligned to the National Science Education Standards and standards for most states. Each module covers an integrated area of science (e.g., Mixtures and Solutions, Measurement, Variables). The instructional materials for each module are packaged in a kit that contains the materials needed to conduct the classroom science investigations: a teacher guide, a module-specific teacher-preparation video, and a summative assessment (Assessing Science Knowledge [ASK]) to be administered before and after each science module.

Within a science module, students in classrooms work in small groups to conduct a series of approximately 16 science investigations over an 8- to 10-week period. These hands-on investigations are aligned to specific science concepts and learning objectives. The structure of the FOSS program provides an ideal test bed for research and evaluation of MyST, with MyST dialogs being aligned with specific classroom science investigations, learning objectives, science standards, and ASK assessments.

Research Motivating the Design of MyST Dialogs

MyST is an example of a new generation of intelligent tutoring systems that facilitate learning through natural spoken dialogs with a virtual tutor in multimedia activities. Intelligent tutoring systems aim to enhance learning achievement by providing students with individualized and adaptive instruction similar to that provided by a knowledgeable human tutor. These systems support typed or spoken input, with the system presenting prompts and feedback via text, a human voice, or an animated pedagogical agent (Graesser, VanLehn, Rosé, Jordan, & Harter, 2001; Lester et al., 1997; Mostow & Aist, 2001; VanLehn et al., 2007; Wise et al., 2005). Text, illustrations, and animations may be incorporated into the dialogs. Research studies show up to one sigma gains (approximately equivalent to an improvement of one letter grade) when comparing performance of high school and college students who use the tutoring systems with students who receive classroom instruction on the same content (Graesser et al., 2001; VanLehn & Graesser, 2001; VanLehn et al., 2005). In a recent synthesis of research that compared learning gains following human tutoring or following use of an intelligent tutoring system, VanLehn (2011) concluded that human tutoring and intelligent tutoring systems produce approximately the same effect size, with human tutoring at $d = 0.79$ and intelligent tutoring systems at $d = 0.76$.

The development of MyST is informed by several decades of research in psychology and computer science. In the remainder of this section, we briefly describe theory and research that informed the design of MyST.

Benefits of Tutorial Instruction

Theory and research provide strong guidelines for designing effective tutoring dialogs. Over two decades of research have demonstrated that learning is most effective when students receive individualized instruction in small groups or one-on-one tutoring. Bloom (1984) determined that the difference between the amount and quality of learning for students who received classroom instruction and those who received either one-on-one or small-group tutoring was two standard deviations. Evidence that tutoring works has been obtained from dozens of well-designed research studies, meta-analyses of research studies (Cohen, Kulik, & Kulik, 1982), and positive outcomes obtained in large-scale tutoring programs (Madden & Slavin, 1989; Topping & Whiteley, 1990).

Benefits of tutoring can be attributed to several factors, including the following:

Question generation. A significant body of research shows that learning improves when teachers and students ask deep-level-reasoning questions (Bloom, 1956). Asking authentic questions leads to improved comprehension, learning, and retention of texts and lectures by college students (Craig, Gholson, Ventura, & Graesser, 2000; Driscoll et al., 2003; King, 1989) and school children (King, 1994; King, Staffieri, & Adelgais, 1998; Palinscar & Brown, 1984).

Generating explanations. Research has demonstrated that having students produce explanations improves learning (Chi et al., 1989; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; King, 1994; King et al., 1998; Palinscar & Brown, 1984). In a series of studies, Chi et al. (1989, 2001) 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, De Leeuw, Chiu, & LaVancher, 1994). Hausmann and Van Lehn (2007a, 2007b) note that "self-explaining has consistently been shown to be effective in producing robust learning gains in the laboratory and in the classroom" (2007b, p. 1067.) Experiments by Hausmann and Van Lehn (2007b) indicate that it is the process of actively producing explanations, rather than the accuracy of the explanations, that makes the biggest contribution to learning.

Knowledge coconstruction. Students coconstruct knowledge when they are provided with the opportunity to express their ideas and to evaluate their thoughts in terms of ideas presented by others. There is compelling evidence that engaging students in meaningful conversations improves learning (Butcher, 2006; Chi et al., 1989; King, 1994; King et al., 1998; Murphy et al., 2009; Palinscar & Brown, 1984; Pine & Messer, 2000; Soter et al., 2008).

Social Constructivism

In social constructivism, learning is viewed as an active social process of constructing knowledge "that occurs through processes of interaction, negotiation, and collaboration" (Palincsar, 1998, p. 365). Vygotsky (1978) stressed the critical role of social interaction within one's culture in acquiring the social and linguistic tools that are the basis of knowledge acquisition. "Learning awakens a variety of internal developmental processes that are able to operate only when the child is interacting with people in his environment" (Vygotsky, 1978, pp. 89 –90). He stressed the importance of having students learn by presenting problems that enable them to scaffold existing knowledge to acquire new knowledge. Vygotsky introduced the concept of the zone of proximal development, "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, p. 86). Social constructivism provides the conceptual model for knowledge acquisition in MyST: to improve learning by scaffolding conversations using open-ended questions and media to support hypothesis generation and coconstruction of knowledge.

Discourse Comprehension Theory

Cognitive learning theorists generally agree that learning occurs most effectively when students are actively engaged in critical thinking and reasoning processes that cause new information to be integrated with prior knowledge. Discourse comprehension theory (Kintsch, 1988, 1998) holds that deep learning requires integration of prior knowledge with new information and results in the ability to use this information constructively in new contexts. To the extent possible, MyST attempts to determine relevant information that students know and build on that lead students to correct explanations.

Social Agency and Pedagogical Agents

When human computer interfaces are consistent with the social conventions that guide our daily interactions with other people, they provide more engaging, satisfying, and effective user experiences (Nass & Brave, 2005; Reeves & Nass, 1996). Such programs foster social agency, enabling users to interact with them the way they interact with people. In comparisons of programs with and without talking heads or human voices, children learned more and reported more satisfaction using programs that incorporated virtual humans (Atkinson, 2002; Baylor & Kim, 2005; Moreno, Mayer, Spires, & Lester, 2001). A number of researchers have observed that children become highly engaged with virtual tutors and appear to interact with a virtual tutor as if it were a real teacher and appear motivated to work hard to please it. Lester (Lester et al., 1997) termed this phenomenon the "persona effect."

Multimedia Learning

During MyST dialogs, students are encouraged to construct explanations of science presented in illustrations, silent animations, and interactive simulations. The design of these dialogs is consistent with research indicating that combining spoken explanations with media can optimize science learning, either during multimedia presentations (Horz & Schnotz, 2010; Mayer, 2001, 2005) or when students are required to generate explanations in multimedia learning environments (Roy & Chi, in press). In a series of studies, Mayer (2001) investigated students' ability to learn how things work (motors, brakes, pumps, lightning) when information was presented in different modalities (e.g., text with illustrations, or narration of the text during which a spoken voice explained the information presented in an illustration or sequence of illustrations). A key finding of Mayer's work is that simultaneously presenting speech (narration) with nonverbal visual information (a sequence of illustrations or an animation) results in the highest retention of information and the application of knowledge to new problems. Mayer (2001) argued that when a person is presented with a well-designed narrated animation, the listener is able to construct an enriched multimodal representation of the two sources of input, leading to superior recall and transfer of knowledge to new tasks. Roy and Chi (in press), based on a review of the literature on self-explanations in multimedia environments, suggest that

> many learners would benefit from self-explanation training or prompting within multimedia environments. Essentially, we have argued that because they are information rich, multimedia environments afford the generation of many opportunities for explaining encoded information and accessing and relating prior knowledge. (p. 27)

Dialog Interaction

The design of spoken dialogs in MyST is based on a number of principles used in Questioning the Author (QtA), an approach to classroom discussions developed by Isabel Beck and Margaret McKeown (Beck, McKeown, Sandora, Kucan, & Worthy, 1996; McKeown & Beck, 1999; McKeown, Beck, Hamilton, & Kucan, 1999). During the 3-year period in which MyST dialogs were designed, tested, and refined, we worked with QtA codeveloper Margaret McKeown to apply principles of QtA to spoken dialogs with Marni that incorporate illustrations, animations, and interactive simulations to help students visualize the science they are trying to explain.

QtA is a mature, scientifically based, and effective program used by hundreds of teachers across the United States. It is designed to improve comprehension of narrative or expository texts that are discussed as they are read aloud in the classroom. The focus is to have students grapple with, and reflect on, what an author is trying to say in order to build a representation from the text. The approach uses open-ended questions to initiate discussion (What is the author trying to say?) to help students focus on the author's message (That's what she says, but what does she mean?) to help students link information (How does that fit with what the author already told us?) and to help the teacher guide students toward comprehension of the text.

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

Murphy and Edwards (2005) analyzed the results of research studies that met rigorous scientific criteria for evaluating programs designed to improve student learning through classroom conversations. Of the nine programs that met the scientific criteria for valid research studies, QtA was identified as one of two approaches that is likely to promote high-level thinking and comprehension of text (Murphy & Edwards, 2005). Moreover, analysis of the QtA discourse showed a relatively high incidence of authentic questions, uptake, and teacher questions that promoted high-level thinking—all indicators of productive discussions likely to promote learning and comprehension of text (Soter & Rudge, 2005).

The MyST System

System Description

Students learn science in MyST through natural spoken dialogs with the virtual tutor Marni, a 3-D computer character that is on screen at all times. Marni asks students open-ended questions related to illustrations, silent animations, or interactive simulations displayed on the computer screen. Figure 1 displays a screen shot of Marni asking questions about media displayed in a tutorial. The student's computer shows a full screen window that contains Marni, a display area for presenting media, and a display button that indicates the listening status of the system. Marni produces accurate visual speech, with head and face movements that are synchronized with her speech.

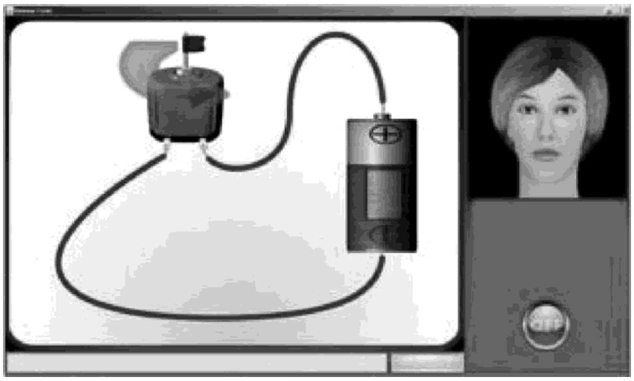


Figure 1. My Science Tutor (MyST) screen layout.

We call these conversations with Marni *multimedia dialogs*, because students simultaneously listen to and think about Marni's questions while viewing illustrations and animations or interacting with a simulation. The media facilitate dialogs with Marni by helping students visualize the science they are discussing. The primary focus of each dialog is to elicit self-explanations from students. MyST analyzes the spoken explanations to determine what the student does and does not know about the science, then presents follow-up questions, which may be accompanied by new media, to help the student construct a correct explanation of the phenomena being studied. The virtual tutor Marni, who speaks with a recorded human voice, is designed to behave like an effective human tutor that the student can relate to and work with to learn science. This is achieved by modeling dialogs between students and human tutors trained in using QtA during the development phase of the project. These dialogs scaffold learning by providing students with support when needed until they can apply new skills and knowledge independently (Vygotsky, 1978).

Marni elicits self-explanations from students using strategies that embody QtA dialog moves such as *marking* and *revoicing*. These two techniques require that the system identify the student's dialog content (marking it) followed by repeating (revoicing) a paraphrase of the information back to the student as a part of the next question: *You mentioned that electricity flows in a closed path. What else can you tell me about how electricity flows?* Marni's responses are designed to communicate this understanding back to the students and to engage and assure them that she understands what they are saying.

A tutorial session generally begins with relating the session to what the student has recently covered in class (during a science investigation), with Marni saying something like: *What have you been studying in science recently?* If the student says something recognizable as the tutorial topic (e.g., "We made a circuit"), the system moves forward by asking the student what they know about the topic: *You mentioned circuits. Can you tell me what a circuit is?* If nothing from what the system extracted from the student's answer relates to the topic, then Marni introduces the topic: *I heard you were learning about circuits. Can you tell me what a circuit is?* For each key concept discussed, the interaction typically begins with a general open-ended question (accompanied by media, such as a picture of a simple circuit): *What's this all about?* or *What's going on here?* and then proceeds to more directed open-ended questions like: Can you tell me more about the flow of electricity in the circuit?

Media are used to ground the conversation, focus the student's attention, help the student visualize the science, and provide a visual frame of reference for the student to talk about. The media are not narrated, and they do not explain the concept to the student. A typical strategy used by MyST is to show an animation to the student and ask him or her to explain what is going on. The use of media was initially intended as a mechanism to get students past *sticking points*, points in a dialog when the system is not able to elicit information from the student that it can build on. During dialogs with project tutors during system development, discussed below, the method proved so useful for eliciting explanations that tutors began to use this as the standard introduction to concepts: ask an introductory question about what a student knows, show an illustration, and ask what is going on.

As noted, MyST dialogs incorporate three types of media: (a) illustrations, (b) animations, and (c) interactive simulations, illustrated in Figure 2. Although these sometimes overlap in the content presented, each plays a unique role. Illustrations are static Flash drawings and are a good way to initiate discussions about topics. They provide the student with a visual frame of reference that helps focus the student's attention and the subsequent discussion on the content of the illustration: *So, what's going on here?* Animations are noninteractive, silent Flash animations that help students visualize concepts that can be difficult to capture in illustrations. In Figure 2, the direction of the flow of electricity is represented by blue dots moving from the D-cell through the wires and bulb and back to the D-cell. The animations enable Marni to ask the student questions to elicit explanations about what is being shown. Simulations allow students to interact directly with the Flash animation using a mouse. Figure 2 shows a simulation of a FOSS classroom investigation called "Breaking the Force" in which students investigate how much weight (number of metal washers) is required on one side of a balance scale to break the force of the magnets attracting each other on the other side. The number of washers in the cup and the space between magnets can be investigated and graphed in this simulation. During multimedia dialogs, as students are interacting with a simulation, the tutor can say things like: *What could you do to . . .? What happens if you . . .?*

System Operation (How Spoken Dialogs Work)

MyST uses character animation, automatic speech recognition, natural language processing, and dialog modeling to support conversations with Marni. The dialogs are designed to elicit responses from students that show their understanding of a specific set of points. The key points of a dialog are specified as propositions realized as semantic frames. The frames represent the events and entities in the domain and the roles that they play. For example, *Current goes from the negative terminal to the positive* would be represented as: **Electricity Flows Origin.negative Destination.positive**. During spoken dialogs, the tutor asks questions that are designed to elicit student responses that will map to the elements of the targeted semantic frames. Information extracted from student responses is integrated into the session context that represents which points have been addressed by the student, which have not, which were expressed correctly, and which represented misconceptions. In analyzing a student's answer, the system tests whether the correct values are filling the semantic roles (i.e., whether the value of Origin is negative or positive). On the basis of the current context, the system generates questions to elicit explanations of the elements needed to produce a complete explanation. Follow-up questions and media presentations are designed to scaffold learning by providing hints about the important elements of the investigation that the student did not include or misunderstood. When possible, the follow-up questions are created by taking a relevant part of the student's response and asking for elaboration, explanation, or connections to other ideas.

This interaction style is well suited to automatic speech recognition (ASR) technology, which will have some amount of recognition error. In sessions in which the system is able to accurately recognize and parse student responses, it is able to adapt the tutorial to the individual student. It may move on to another point or delve more deeply into a discussion of concepts that were not correctly expressed by the student, using marking and revoicing to incorporate information from the student's response. If the student does not seem to grasp the basic elements under discussion, the system presents more background material. If the system is unable to elicit and understand relevant student responses, by default it proceeds through the session with a full discussion of each point.

Using spoken responses in this way can increase efficiency and naturalness of the interaction while minimizing the impact of system errors. False-negative errors, in which the system does not recognize correct information provided by the student, simply cause the system to continue 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, may cause the system to move on from a point before it is sufficiently

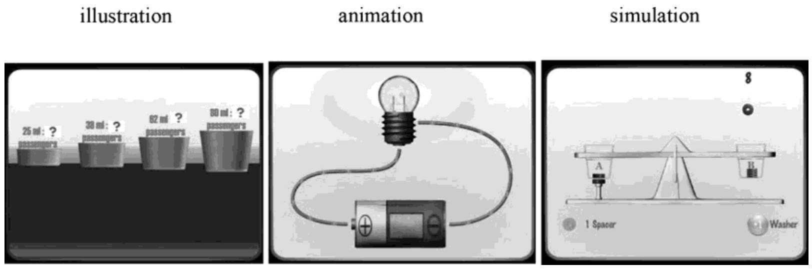


Figure 2. Media types.

covered. False-accept errors are rare and have not proved to be a problem.

System Development

During the development and evaluation of MyST, data were collected from tutoring sessions at elementary schools in the Boulder Valley School District (BVSD). A team of project tutors was trained in the FOSS content and QtA-based interaction style. Using FOSS teacher guides, the team developed learning objectives and specifications for media presentations aligned to each classroom science investigation. Tutors went into the schools and tutored students using the materials developed. Visuals were presented on laptops, and students wore headsets for recording their speech. The recorded sessions were reviewed in group meetings to revise the presentations and determine *sticking points* that would benefit from the introduction of media. These meetings also helped foster a common style across tutors. In addition, transcripts of tutoring sessions were reviewed and annotated by M. McKeown to provide constructive feedback to the project tutors on how to use QtA principles most effectively. The data collected in the human-tutored sessions were used to train the speech recognition and natural language-processing modules to interpret the students' speech and to develop dialog models to attempt to emulate the behavior of the human tutors. These modules were integrated to produce the first version of MyST that was used in Wizard-of-OZ (WOZ) studies.

WOZ

WOZ data collection attempts to provide user interactions similar to the target application, but a human controls the system behavior. In the WOZ collection, students independently interacted with Marni, while a remote human tutor, connected to the student's computer via the Internet, monitored and controlled the system's behavior. The human wizard could see everything on the student's computer and hear what the student was saying. At each point in a dialog when the system was about to take an action (e.g., have Marni talk; present a new illustration), the action was first shown to the human wizard who could accept or change the action. The system logged all transactions during the session. Transcriptions of the dialogs in each session were then reviewed by developers to refine the dialog model. The primary changes during this phase of development included adding new media, expanding the coverage of the natural language processing (to accommodate new ways students could talk about concepts), and adding new ways of asking students questions. As the tutorials evolved, human wizards intervened less.

In sum, during initial development of tutorial dialogs with human tutors, a total of 189 students received human tutoring over a total of 427 sessions. During the subsequent WOZ sessions, a total of 347 students received WOZ tutoring over 1,156 sessions. The purpose of data collected during development was to improve system coverage, that is, modeling the different ways that diverse students talked about science and refine the media presentations, so the emphasis was on including a greater variety of students, with less data from each individual student than in the system evaluation.

System Evaluation

All data collected in the human-tutoring and WOZ sessions were used to train the final acoustic, language, and dialog models for the virtual tutoring system. During the 2010 –2011 school year, an assessment of the MyST system was conducted to examine the effect of the virtual tutor on student test scores in science. During the assessment, students interacted with Marni independently in their schools, without a human wizard. An experimenter logged students into the MyST system and specified the dialog session to be used, but otherwise left students alone to use the system. The experimental design compared students receiving MyST tutoring with those receiving face-to-face human tutoring in small groups. Students were randomly assigned within classrooms to tutoring condition, and these groups were also compared with students from intact control classrooms with no tutoring. Students completed one of four FOSS modules (*Variables*, *Magnetism*, and *Electricity*, *Measurement and Water*) and were tested pre–post with the FOSS-ASK assessment for that module. All students received similar classroom instruction. The two hypotheses for the study were as follows:

> *Hypothesis 1:* Students receiving tutoring with MyST will show learning gains roughly similar to students receiving face-to-face human tutoring.
>
> *Hypothesis 2:* Both groups receiving tutoring will show greater learning gains than students receiving no tutoring.

Method

Participants

Data were collected from tutoring sessions at elementary schools in the BVSD. BVSD is a 27,000-student school district with 34 elementary schools. There is substantial student diversity across schools, which vary from low to high performing on state science tests. A list of potential schools was developed in collaboration with the BVSD science director. All third-, fourth-, and fifth-grade teachers at these schools were invited to participate in the study, and teachers who accepted were enrolled in the study. All students in the classrooms of participating teachers were invited to participate. All students who agreed to participate were enrolled. All third-, fourth-, and fifth-grade teachers in the district who did not participate as treatment classrooms were recruited to serve as control class- rooms, and those who agreed were enrolled.

The data set contained 1,478 students at 22 schools and 63 classrooms. One hundred two students in 14 classrooms in six schools were tutored with MyST, and 85 students in these same classrooms received human tutoring. Control students accounted for 1,155 students in 49 classrooms and 19 schools. These students received no tutoring, but did receive instruction in FOSS modules during class. For analysis, nonconsented students were removed from the sample. Other reasons for removing students from the sample included unmatched pre–post tests where students did not fill out a majority of answers and tests with grading concerns, including very low reliabilities. The remaining sample totaled 1,167 students. Eighty-three students received MyST tutoring, 69 were tutored in small groups (both in 12 classrooms), and 1,015 students in 50 classrooms in 20 schools received only classroom instruction and no tutoring. All missing data were removed by an analyst who was blind to the experimental condition.

Procedure

Consented students in the study were assigned to receive tutoring *in addition to* their normal classroom instruction for the module. Teachers specified the space in the school to be used, and this varied from school to school, generally any relatively quiet room. The teacher also scheduled the time for their students to minimize the impact on the student's other activities. Tutoring times were always during regular school hours. General guidelines were that this time should not be at recess or lunch, during core subject time (reading, math, science), or during special activities time (art, music).

All students in the study received in-class instruction in the FOSS modules: Measurement (third grade), Magnetism and Electricity (fourth grade), Water (fourth grade), and Variables (fifth grade). Teachers in both treatment and control classrooms followed module lesson plans and used FOSS materials. Students participating in the study received tutoring from MyST or human tutors for 12–16 20-min sessions concurrent with their regular classroom instruction. Each tutorial was oriented around a set of key concepts the student was expected to have learned from classroom instructional activities. Both MyST and human tutoring used the same multimedia content linked to FOSS content. MyST students were tutored individually on computers. Headsets with earphones and microphones were used to reduce noise interference. For most sessions, eight students at a time used the computers in a separate resource room at each school. Students in the human tutoring condition received tutoring with human tutors for the same amount of time as those in the MyST group. They worked in groups of three to four students with each human tutor. Although one-on-one interaction with a human tutor would present a more direct comparison to the virtual tutor condition, the study did not have sufficient resources to provide one-on-one human tutoring; however, research has demonstrated equivalent learning gains for one-on-one and small-group tutoring (e.g., Bloom, 1984).

Measures

Students in all experimental groups were given the ASK summative assessments as pre- and posttest measures. Tests were administered before the beginning of the FOSS lessons for the module, and immediately after tutoring for the module ended. The ASK assessments for the four modules used in the assessment have identical pre and post versions. Depending on the module, the assessments have between eight and 12 items, consisting of multiple-choice and constructed response questions, and show composite internal reliability with alphas in the range of 0.80–0.90. The interrater reliability for subjective items has also met high standards in similar conditions (e.g., *r* = .90), and the validity of the measures has been built up over time through a process of empirical investigation.

Because module tests have different scales, scores were standardized to a common metric. All standardization was conducted on data with outliers and other spurious data removed. "Testwise" standardization subtracted the mean of each test (over all students and pooling pre/post) from each student's score. This difference was then divided by the average standard deviation for both pre and post for each test.

Pairs of raters (tutors) scored all assessments from tutored students and a subset of assessments from control students. Raters trained together with scoring rubrics provided by FOSS, then scored the assessments independently. All scoring was blind to experimental condition (human tutor, virtual tutor, no tutoring) and whether the assessment was pre or post. Interrater reliabilities for two raters were high (counting only the open-ended items), with intraclass correlation coefficients ranging from .89 to .98, with averages for pre and post of .93 and .94, respectively. Internal reliabilities (Cronbach's alpha) were lower, ranging from a = .60 to a = .89 for both pre and post versions of the assessments, with averages for pre = .74 and post = .79. Scores used for outcome analysis were the averages across both raters.

Results

Several comparisons were made to test the hypotheses. To make comparisons, both standardized pre/post scores and *residual gain scores* compared groups on the average differences between their observed and expected scores. Gain differed markedly depending on where students started on the pretest, regardless of which group they belonged to. Students who started lower on the pretest gained more than students starting higher. This is often a sign of regression toward the mean where greater gain occurs for students starting lower regardless of actual learning. Regression toward the mean complicated the group comparisons for this study because the control students on average scored much lower on the pretest than students receiving tutoring. We believe the lower pretest scores for the control were primarily due to two factors:

1. Consented students (those whose parents returned signed permission forms) had higher pretest scores than nonconsented students. Pretest scores for nonconsented students were similar to the control group.

2. Schools choosing to participate as treatment groups in the study were not representative of the overall free and reduced lunch (FRL) percentage of the district. Boulder Language Technologies worked with BVSD officials to identify a set of schools to recruit. All classroom teachers for the targeted grades in those schools were recruited, and all of the teachers who agreed to participate were enrolled. In this particular study, those teachers who agreed to participate represented schools that had smaller percentages of FRL students. Schools with higher percentages of FRL students tend to have lower test scores, and more of these schools were in the control group.

When group comparisons were made, control students tended to gain more pre to post than tutored students simply because they started lower on the pretest. Residual gain scores and analysis of covariance (ANCOVA) were used for analysis to adjust for these differences in prescore (Rudestam & Newton, 1999). The residual gain score is the observed score minus the expected score in the scatter between pre and post; the expected score is the regression line for the scatter. It is used to compare groups and has a mean of zero, with a scale representing standard deviation units.

Comparison Between Tutored Groups

The first hypothesis examined whether MyST and human-tutored groups were roughly equal to each other in pre/post gain. Students were randomly assigned within classrooms to tutoring conditions. Standardized gain for the human-tutored group ($M =$ 1.95, SD = 0.85) was not significantly different than for the MyST-tutored group ($M = 1.75$, $SD = 1.03$), $t(150) = -1.31$, $p =$ .190, d = .18. Residual gain for the human-tutored group ($M =$ 0.51, SD = 0.66) was also not significantly different than for the MyST-tutored group ($M = 0.38$, $SD = 0.76$), $t(150) = -1.15$, $p =$ .250, d = .15. Power analysis showed that for an effect size of $d =$ .15, sample sizes of 600 students per group would be needed to reach significance at the .05 level with 80% power. The small effect size and lack of statistical significance support the first hypothesis that benefits of tutoring are roughly equal for human tutors and Marni in pre/post gain.

Comparison With Control Group

As stated, comparisons with the students in control classrooms were complicated by differences in pre-test scores. To adjust for these differences, comparisons were made with residual gain scores and an ANCOVA to test the second hypothesis that students in tutored groups gained more than students in the control group. Standardized gain scores showed a moderate difference between MyST (M = 1.75, SD = 1.03) and control (M = 1.57, SD = 1.01; d = .18) and a larger difference between the human (M = 1.95, SD = 0.86) and control (d = .40). Effect sizes for residual gain scores were calculated by the difference in means between groups divided by the pooled standard deviation for the residual gain distribution. A moderate effect size was observed for the comparison of MyST tutoring (M = .38, SD = .76) and control ($M =$ −.06, SD = .84; d = 0.53) and a larger effect size for human tutoring (M = .51, SD = .66) and control (d = 0.68). A one-way analysis of variance (ANOVA) tested whether group means differed significantly on residual gain score. The main effect for tutoring was significant, $F(2, 1164) = 26.06$, $p < .001$. Post hoc tests showed significant differences between both tutoring groups and the control group, and no significant differences between the two tutoring groups.

An ANCOVA confirmed the findings from the analysis of residual gains. Like residual gain scores, ANCOVA also adjusts group means for differences in pretest. ANCOVA in this context gave almost identical results to the ANOVA using residual gains, $F(2, 1163) = 26.60$, $p < .001$. Comparisons of adjusted means were also nearly identical to effect sizes in residual gains for groups. ANOVA and ANCOVA tests support the second hypothesis that tutored groups gain significantly more from pre to post than students in the control group.

Gain was also assessed as a function of prescore. Group comparisons divided the prescore distribution for the tutored group into five equal parts. All groups showed higher gain for the lower prescore blocks.

The use of hierarchical models allows for partitioning of error between students and classrooms, and quantifying how much total variability is due to each level. Estimates of classroom variability, calculated with all students in the classroom, equaled 46%. Hypothesis testing for classroom effects showed significant effects for both MyST compared with control, $t(60) = 2.5$, $p = .014$, and human compared with control, $t(60) = 3.0$, $p = .004$. These results from hierarchical models also support the second hypothesis that tutored groups gain more from pre to post than the control group.

Component Evaluation

In order to evaluate the performance of the speech-processing components, student utterances for a subset of the assessment data were manually transcribed and parsed into frames to give the reference data to compare against. ASR performance is typically expressed as a word error rate (WER), which is the sum of word deletion, insertion, and substitution errors divided by the number of words in the reference string (from human transcriptions). The speech recognizer vocabulary size was 6,235 words. The WER for the assessment sessions was 41.4%.[1] This is a large WER, and would not be viable for many applications. The system performed well even with the high WER because the accuracy of extraction of frame elements (the key concepts being discussed) from student's speech remained relatively high, with an overall Recall = 79% and Precision = 82%. So 79% of the relevant information in the reference parses was correctly extracted from the ASR output. Of the information extracted, 82% of the elements were correct. These results indicate that many of the recognition errors were in information that was not relevant or redundant. Given the nature of QtA dialogs and the way spoken responses are used by the system, this level of extraction accuracy was sufficient to produce both engaging and effective dialogs, as indicated by students' responses to questionnaires and the learning gains.

Survey Results

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 the following: (a) Written (vs. oral) surveys for students were administered, (b) students were verbally assured of anonymity, (c) questionnaires were anonymous in that students did not write their names on the survey, and (d) 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. 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

[1] The performance of the ASR system was enhanced significantly over the course of the project, and WER on the assessment data is now 21%. However, the system and models were fixed at the start of the assessment to avoid confounding the evaluation results with improvements in the performance of the speech recognition system.

53% selected “I am more excited about science” after using the program. Only 4% felt that the tutoring did not help.

Teachers were asked for feedback to help assess the feasibility of an intervention using the system and their perceptions of the impact of the system. A teacher survey was given 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 nine 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. Of the responding teachers, 100% 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. In addition, 93% said that they would like to participate in the project again. Furthermore, 74% 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.

Conclusion

In the present article, we presented the motivation, design, and evaluation results for a conversational multimedia virtual tutor for elementary school science. The operating principles for the tutor are grounded in research from education and cognitive science. Speech, language, and character animation technologies play a central role because the focus of the system is on engagement and spoken explanations by students during spoken dialogs with a virtual tutor.

An assessment was conducted in schools to compare learning gains from human tutoring and MyST with business-as-usual classrooms. Both tutoring conditions had significantly higher learning gains than the control group. Although the effect size for human tutors versus control ($d = 0.68$) was larger than for MyST versus control ($d = 0.53$), statistical tests supported the hypothesis of no significant difference between the two.

After the assessment, surveys were collected from students and teachers that bear on the engagement and feasibility of the tutoring system. Following a series of tutoring sessions with Marni, the great majority of students reported that they enjoyed spending time working with her, that they felt that Marni helped them learn science, and that they felt more interested in science and more motivated to learn science than they had before using the system. Teachers reported that they would like to use MyST in the future to tutor all of their students and that they would recommend the program to other teachers.

One conclusion that we draw from this study is that current spoken dialog and character animation technologies can be combined with media to provide engaging and effective experiences for third-, fourth-, and fifth-grade students learning science. Students who used MyST interacted with Marni for 4 –5 hr over the course of the 16 dialog sessions over an 8- to 10-week period. No students dropped out of the study, and the large majority of students reported positive experiences. We believe that the QtA approach helped assure the student that Marni is listening to and understands what they are saying; this experience is fostered by dialog moves such as revoicing and marking that Marni produces. Dialogs based on QtA enable the tutorial dialog to proceed in a graceful way even when the system does not accurately interpret what the student said, because the system typically proceeds with a reasonable follow-up question, which the student accepts as a natural extension of the dialog.

The system described presents baseline results for one specific system based on a number of design decisions. Further work is needed to understand the effects of the individual features of the system. For example, we do not know the relative contribution of media in helping students visualize science and construct explanations, or the contribution of the dialog moves and questions that Marni generated, to the learning gains that occurred. We believe the MyST system provides a framework and infrastructure for conducting research on these questions. Planned future work will allow us to expand the context of the interaction from one-on-one tutoring to systems that support conversations in which a virtual tutor is able to mediate conversations among small groups of students. The virtual tutor will then be able to ask questions that help students build on each other’s ideas to coconstruct explanations consistent with accurate mental models of the science.

References

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology, 94,* 416–427. doi:10.1037/0022-0663.94.2.416

Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education, 15,* 95–115.

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,* 385– 414. doi: 10.1086/461835

Bloom, B. (1956). *Taxonomy of educational objectives, Handbook I: The cognitive domain*. New York, NY: David McKay.

Bloom, B. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13,* 4 –16.

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology, 98,* 182–197. doi:10.1037/0022-0663.98.1.182

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

Chi, M. (2009). Active– constructive–interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1,* 73–105.

Chi, M., Bassok, M., Lewis, M., Reimann, P., Glaser, R., & Alexander. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145–182. doi: 10.1207/s15516709cog1302_1

Chi, M., De Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439 – 477.

Chi, M., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25,* 471–533. doi: 10.1207/s15516709cog2504_1

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.

Craig, S., Gholson, B., Ventura, M., & Graesser, A. (2000). Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education, 11,* 242–253.

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,* 431– 450. doi:10.2190/Q8CM-FH7L-6HJU-DT9W

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington DC: National Academy Press.

FOSS. (n.d.). *About FOSS*. Retrieved from http://www.fossweb.com

Gamoran, A., & Nystrand, M. (1991). Background and instructional effects on achievement on eighth-grade English and social studies. *Journal of Research on Adolescence, 1,* 277–300.

Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22,* 39 –51.

Hake, R. (1998). Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66,* 64 –74.

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

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

Horz, H., & Schnotz, W. (2010). Multimedia: How to combine language and visuals. *Language at Work—Bridging Theory and Practice*. Retrieved from http://ojs.statsbiblioteket.dk/index.php/law/article/view/6200

King, A. (1989). Effects of self-questioning training on college students' comprehension of lectures. *Contemporary Educational Psychology, 14,* 366 –381. doi:10.1016/0361-476X(89)90022-2

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31,* 338 –368.

King, A., Staffieri, A., & Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology, 90,* 134 –152. doi:10.1037/0022-0663.90.1.134

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163–182. doi:10.1037/0033-295X.95.2.163

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). *The persona effect: Affective impact of animated pedagogical agents*. Paper presented at the Proceedings of the SIGCHI conference on Human Factors in Computing Systems, Atlanta, GA.

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 (pp. 52–72). Boston, MA: Allyn & Bacon.

Mayer, R. (2001). *Multimedia learning*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139164603

Mayer, R. (2005). Introduction to multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 1–16). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511816819.002

McKeown, M., & Beck, I. (1999). Getting the discussion started. *Educational Leadership, 57,* 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.

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction, 19,* 177–213. doi:10.1207/S1532690XCI1902_02

Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart machines in education: The coming revolution in educational technology* (pp. 169 –234). Cambridge, MA: MIT Press.

Murphy, P. K., & Edwards, M. N. (2005). *What the studies tell us: A meta-analysis of discussion approaches*. Paper presented at the American Educational Research Association, Montreal, Canada.

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,* 740 –764. doi:10.1037/a0015576

Nass, C., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. Cambridge, MA: MIT Press.

National Assessment of Educational Progress. (2005). *National and state reports in science: The nation's report card*. Jessup, MD: ED Pubs.

Nystrand, M. (1997). *Opening dialogue: Understanding the dynamics of language and learning in the English classroom*. New York, NY: Teachers College Press.

Osborne, J. (2010, April 23). Arguing to learn in science: The role of collaborative, critical discourse. *Science, 328,* 463– 466.

Palincsar, A. S. (1998). Social constructivist perspectives on teaching and learning. *Annual Review of Psychology, 49,* 345–375. doi:10.1146/annurev.psych.49.1.345

Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1,* 117–175. doi:10.1207/s1532690xci0102_1

Pine, K., & Messer, D. (2000). The effect of explaining another's actions on children's implicit theories of balance. *Cognition and Instruction, 18,* 35–51. doi:10.1207/S1532690XCI1801_02

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

Roy, M., & Chi, M. (in press). The self-explanation principle. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning*.

Rudestam, E., & Newton, R. (1999). *Your statistical consultant: Answers to your data analysis questions*. Washington DC: Sage.

Soter, A. O., & Rudge, L. (2005). *What the discourse tells us: Talk and indicators of high-level comprehension*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

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. doi:10.1016/j.ijer.2009.01.001

Topping, K., & Whiteley, M. (1990). Participant evaluation of parent-tutored and peer-tutored projects in reading. *Educational Research, 32,* 14 –32. doi:10.1080/0013188900320102

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46,* 197–221. doi:10.1080/00461520.2011.611369

VanLehn, K., & Graesser, A. C. (2001). *Why2 Report: Evaluation of Why/Atlas, Why/AutoTutor, and accomplished human tutors on learning gains for qualitative physics problems and explanations*. Unpublished report, University of Pittsburgh CIRCLE group and the University of Memphis Tutoring Research Group.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31,* 3– 62. doi:10.1080/03640210709336984

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J., Shelby, R., Taylor, L., . . . Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education, 15,* 147–204.

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

Wise, B., Cole, R., Van Vuuren, S., Schwartz, S., Snyder, L., Ngampatipatpong, N.,... Pellom, B. (2005). *Learning to read with a virtual tutor: Foundations to literacy: Interactive literacy education: Facilitation literacy environments through technology*. Mahwah, NJ: Erlbaum.

Received December 15, 2011
Revision received November 19, 2012
Accepted December 10, 2012 ■

Developing Conversational Multimedia Tutorial Dialogs

Wayne Ward[1,2] and Ron Cole[1]
[1] Boulder Language Technologies; [2] University of Colorado

Cite as: Ward, W. & Cole, R. (2015) Developing Conversational Multimedia Tutorial Dialogs. In R. Sottilare, A. Graesser, X. Hu & K. Brawner (Eds.), *Design Recommendations for Intelligent Tutoring Systems*, *Vol 3. Authoring Tools and Expert Modeling Techniques.* Chapter 20, pp. 243 – 254. U.S. Army Research Laboratory. file:///C:/Users/Ron/Downloads/Design_Recommendations_for_Intelligent_Tutoring_Systems_Volume3Final%20(2).pdf

INTRODUCTION

This chapter describes an approach to authoring intelligent tutoring systems used in My Science Tutor (MyST). This virtual science tutor engages children in spoken dialogs in which they learn to construct explanations of science phenomena presented in illustrations, animations and interactive simulations. Tutorials are developed through an iterative process of recording, annotating and analyzing logs from sessions with students, and then updating tutor models. This approach has been used to develop over 100 tutorial dialog sessions, of about 15 minutes each, in 8 areas of elementary school science. Summative evaluations indicate that students are highly engaged in the tutoring sessions, and achieve learning outcomes equivalent to expert human tutors (Ward et al., 2011; 2013).

This chapter describes the process of developing conversational science tutors that use visual media and the infrastructure supporting the development. A particular focus is the development of models for representing and extracting the semantics that provide the basis for selecting tutor actions based on interpretations of student answers. While initial evidence suggests that MyST tutorials can improve students' motivation and science learning (Ward et al., 2011; 2013), the potential of these systems to transform learning and education is limited by the amount of effort required to develop them. A major focus of our current research, discussed in this chapter, is to motivate and demonstrate the feasibility of an approach to authoring conversational tutoring systems that substantially reduces the effort and data required to develop dialogs for each new science domain.

RELATED RESEARCH

Research in intelligent tutoring systems addresses a critical need to provide teachers and students with accessible, inexpensive and reliably effective tools for improving young learners' interest in science, as well as their ability to learn science and participate productively in classroom science activities. The 2009 National Assessment of Educational Progress (NAEP 2009) reports that fewer than 2% of 4th, 8th, and 12th grade students demonstrated advanced knowledge of science, and over two-thirds of all students in these grades were scored as *not proficient in science*. Analyses of NAEP scores in reading, math and science over the past twenty years indicate that this situation is not improving and is actually worsening. The gap between English learners and English-only students, which is over one standard deviation lower for English learners, has increased rather than decreased over the past 20 years.

Intelligent tutoring systems aim to enhance learning by providing students with individualized and adaptive instruction similar to that provided by a knowledgeable human tutor. These systems support conversational interaction with users through either typed or spoken input with the system presenting prompts and feedback via text, human voice or an animated pedagogical agent (Graesser et al., 2001; D'Mello et al., 2011; Rus et al., 2013; Graesser et al, 2014). Advances in intelligent tutoring systems during the past 15 years have resulted in systems that produce learning gains equivalent to human tutoring, which is widely regarded as the most efficient and effective form of learning. A review by Van Lehn (2011) compared learning gains with human tutoring and intelligent tutoring systems that required students to engage in problem solving and construct explanations. 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. Van Lehn concluded that intelligent tutoring systems "are nearly as effective as human tutoring systems." (Van Lehn, 2011, pg. 197). A recent meta-analysis by Ma et al. (2014) indicated that intelligent tutoring systems produce significant effects across a wide range of subjects at all education levels relative to large group instruction, non-ITS computer-based instruction, or textbook or workbooks, and no differences between human tutoring and learning using intelligent tutoring systems. [REF: Ma, W., Adesope, O., Nesbit, J., & Liu, Q. (2014) Intelligent tutoring systems and learning outcomes: A meta-analysis. (2014). *Journal of Educational Technology*, *106,* 901-918.

Research in argumentation and collaborative discourse acknowledges the strong influence of the theories of Vygotsky (1978, 1987) and Bakhtin (1975; 1986), who argue that all learning occurs in and is shaped by the social, cultural and linguistic contexts in which they occur. Roth (2013, 2014) provides an excellent integration of Vygotsky's and Bakhtin's theories and their relevance to research on collaborative discourse. He argues that, when considered in the context of the basic tenets of their theories, "currently available analyses of science classroom talk do not appear to exhibit sufficient appreciation of the fact that words, statements, and language are living phenomena, that is, they inherently change inspeaking." (Roth, 2014). Vygotsky argued that scientific vocabulary and concepts could only be learned through deliberate instruction in an academic setting, as opposed to the more ad hoc manner in which vocabulary and concepts are learned in everyday conversation. Consistent with this view, the 2007 NRC report emphasizes 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 (Duschl et al., 2007).

The past decade has seen a remarkable growth in publications investigating scientific discourse and argumentation. Kuhn (2010) notes that argumentation has become widely advocated as a framework for science education. The idea that argumentation has become both a reform movement and framework for science education is supported by growing evidence of substantial benefits of explicit instruction and practice on the quality of students' argumentation and learning (Chin & Osborne, 2010; Kulatunga & Lewis, 2013). Evidence from these studies indicates that argumentation can be improved by providing professional development to teachers or knowledgeable students (Bricker & Bell, 2009; Bricker & Bell, 2014; deJong, 2013; Berland, 2009), explicitly teaching students the structure of good arguments, and providing students with scaffolds during argumentation that helps them provide evidence for their own arguments and critiquing other's arguments (Kulatunga et al., 2013; Kulatunga and Lewis, 2013).

In the remainder of this chapter, the type of interaction used by MyST is described along with the semantic representation used to support the interaction. The process for developing tutorials is explained with a focus on creation and refinement of the model for extracting semantic representations from spoken student responses. A new approach is then presented for developing more robust semantic parsers for the domain with significantly reduced developer effort.

DISCUSSION

The Nature of Tutorial Dialogs between Students and Marni in My Science Tutor

Since 2007, our research has focused on development of My Science Tutor, an intelligent tutoring system designed to improve science learning of 3rd, 4th and 5th grade children through spoken dialogs with Marni, a virtual science tutor. Because many elementary school children have difficulty reading at grade level, we decided to develop tutoring systems in which students use speech to converse with a virtual tutor. Students in our study received eight to ten weeks of classroom instruction in one of four areas of science—Measurement, Water, Magnetism & Electricity, or Variables—using the Full Option Science System (FOSS, 2014). Over the course of each FOSS module instruction, students conducted 16 science investigations in small groups. Students made written entries and drawings in science notebooks about their predictions, observations and explanations of the science encountered in each investigation. Shortly after each investigation, students engaged in spoken dialogs for 15 to 20 minutes with the virtual tutor Marni or with an expert human tutor. In these dialogs, the human or virtual tutors asked open-ended questions about the science encountered in the classroom science investigations. The tutors asked students questions about science presented in illustrations, animations or interactive simulations to scaffold learning and help them construct accurate and complete explanations. Analyses of dialogs indicate that, during a dialog of about 15 minutes, tutors and students produced about the same amount of speech, around 5 minutes each. The main result of the summative evaluation was that, relative to students in classrooms who did not receive supplemental tutoring, students who were tutored by Marni and by human tutors achieved equivalent learning gains, with moderate to strong effect sizes. Surveys indicated that over 70% of students tutored by Marni reported that they were more excited about studying science in the future. Details of these experiments are reported in Ward et al. (2011, 2013).

It is noteworthy that tutoring by both human and virtual tutors produced significant learning gains, relative to students who did not receive tutoring, given that all students in the study received classroom instruction using a highly respected inquiry-based learning program (FOSS, 2014), that is used by over 1 million K-8 students annually in the U.S. These results are consistent with a meta-analysis by Chi (2009) which indicates that students whose instruction involves *interactive tasks* that include collaborative discourse and argumentation learn more than students whose learning involves *constructive tasks*, (e.g., classroom investigations and written reports), or *active tasks* (e.g., classroom Science Investigations). Chi's synthesis of research indicates the critical importance of having students talk about and explain science to optimize learning in inquiry-based programs.

When using MyST, the student's computer shows a full screen window that contains the virtual tutor Marni (a 3D character), a display area for presenting information 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 is recorded by an experienced science tutor, the voice talent whose phrasing and prosody imbues Marni with the personality of a sensitive and supportive tutor. Spoken dialogs involve Marni asking open-ended questions about science presented in illustrations, silent animations and interactive simulations. Interactive simulations allow students to use a mouse to manipulate variables and observe the effects, such as adding additional winds of wire to an electromagnet core and observing the effect on the number of washers picked up. The pedagogical role of these media types are discussed in detail in Ward et al. (2011). Figure 1 shows a screen shot of the student's screen for the example interactive.

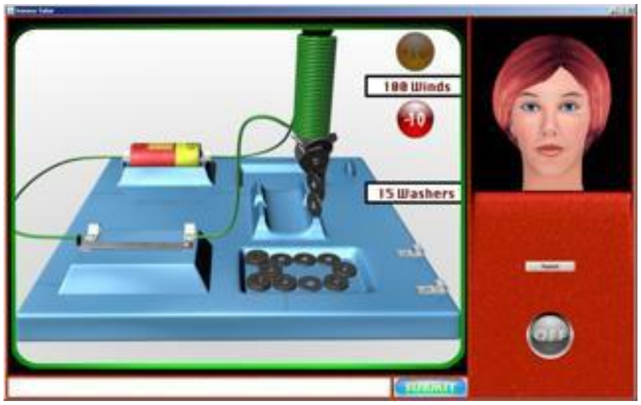


Figure 1: The student screen contains the avatar Marni, a display area, and a listening indicator.

A typical sequence of actions for the tutor 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. Students wear high quality headphones with a noise-cancelling microphone. When ready to speak, the student holds down the space bar. As the student speaks, the audio data are sent to the speech recognition system. When the space bar is released, the word string produced by the speech recognizer is parsed to produce a set of semantic parses. The set of parses is pruned using session context information to a single best interpretation., The new information is added to the session context and a new set of tutor actions is generated. The actions are executed and the system again waits for a student response.

The focus of the MyST system is to elicit explanations of science concepts from students. Each 15 to 20 minute MyST dialog session functions as an independent learning activity that provides, to the extent possible, the scaffolding required to stimulate students to think, reason and talk about science during spoken dialogs with the virtual tutor. The goal of these *multimedia dialogs* is to help students construct explanations that express their ideas. The dialogs are designed so that over the course of the conversation with Marni, the student is able to 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. 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 is to decide when students need to be provided with specific information (e.g., a narrated animation) 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, such as brief multimedia presentation that explains the key concepts the student was attempting to explain.

MyST tutorials are characterized by two key features: the inclusion of media throughout the dialog, and the use of open-ended questions related to the phenomena and concepts presented via the media. Follow-on questions attempt to build on things the student said. For example, an initial classroom investigation about magnets has students move around the classroom exploring and writing down what things do and do not stick to their magnets. The subsequent multimedia dialog with Marni begins with an animation that shows a magnet being moved over a set of identifiable objects, which picks up some of the objects but not others. Marni then says: “What’s going on here?” If the student says: “The magnet picked up some of the objects,” Marni might say: “Tell me more about the types of objects magnets pick up.”

Each tutorial session in MyST is designed to cover a few main points (typically 2 to 4) in a 15 to 20-minute session with a student. The tutorial dialog is designed to get students to articulate concepts and be able to explain processes underlying their thinking. Tutor actions are designed to encourage students to share what they know and help them articulate why they know what they know. For the system (Marni), the goal of a tutorial session is 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 student. Many dialog moves are adapted from principles of Questioning the Author (QtA) (Beck & McKeown, 2006). Much use is made of open-end questions such as "What do you think is going on here?" One of the developers of QtA, Margaret McKeown, worked closely with our development team during development of MyST dialogs. Dr. McKeown analyzed annotations of sessions with human tutors trained in QtA dialog moves, and provided feedback that were used to improve subsequent dialogs. Analysis of MyST Dialogs (Ward et al., 2011; 2013) reveals that concepts expressed by students are recognized at about 85% accuracy. The system fails to recognize about 15% of the concepts correctly expressed by the student. MyST does not tell students that they are wrong, but simply moves on to other propositions if the student expressed understanding, or continues to discuss the current topic otherwise. This strategy provides for graceful dialogs when concept recognition errors occur.

Semantic Representation

The MyST dialog model is based on representing what students are saying about attributes of entities and how entities and events in the domain are related. MyST uses the Phoenix system for Natural Language Processing and for generating tutor moves. Phoenix represents the propositions being discussed as semantic frames with role labels similar to other semantic parsing systems such as FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005), but uses role labels specific to the domain of Science. Roles represent how entities are related to each other and to predicates (usually a verb or nominalization). Semantic frames are used to represent role sets important for the domain. The following example of a statement describing movement would be extracted as:

Electricity flows from the negative terminal through the bulb and to the positive terminal
  Frame: DescribeMovement
  Predicate: Move
  Theme: Electricity
  Source: Terminal.negative
  Goal: Terminal.positive
  Path: Bulb

Other examples of frames important in science discourse are:

Grass is a producer
  Frame: ClassMembership
  Member: Grass
  Class: Producer

The bulbs are not shining because the pathway for electricity to flow has been broken
  Frame: CausalRelation
  Result:
    Theme: Bulb
    State: Off
  Cause:
    Predicate: Interrupted
    Theme: Pathway

Student responses are extracted by the system into semantic frames. Tutor next moves are selected by comparing the frames extracted from student responses to reference frames representing correct role assignments. The following sections explain how role extraction is accomplished and how the extracted frames are used in generating tutor moves.

Defining and Extracting Semantic Frames

The first step in developing a MyST tutorial dialog is to define the topics to be covered. The specification of tutorial semantics begins with creating a narrative. The tutorial narrative is a set of natural language statements that express the concepts to be discussed in as simple a form as possible. These do not represent the questions that the system asks, but are the set of points that the student should express. The narrative represents what an ideal explanation from a student would look like. The narrative statements are manually annotated to reflect the desired semantic parse. An example annotation is:

The current flows from the minus terminal to the plus
Theme: [Electricity] (The current)
Predicate: [Move] (flows)
Source: from the [_negative] (minus terminal)
Goal: to the [_positive] (plus)

Which results in the extracted frame:

Theme: Electricity
Predicate: Move
Source: negative
Goal: positive

These parsed statements define the domain of the tutorial. After enumerating the concepts to be discussed, the visuals to be used to illustrate scientific vocabulary, materials and phenomena and are defined. A short narrative is written and parsed for each of the media files to be used in the tutorial. The Phoenix compiler is used to compile the annotated narratives into Recursive Transition Networks that are used by the parser to extract text into semantic frames.

Student responses are also parsed into the same semantic representations as the narratives. The initial patterns are created from the narratives and have all of the roles and entities that will be discussed, but only a few ways of expressing them. Over the course of development, the patterns must be expanded to cover the various ways students articulate their understandings of the science concepts. In developing the MyST system, project tutors were asked to type simulated student input. These inputs were annotated and added to the training data for the extraction patterns. Once the initial components for a tutorial have been specified, the task becomes to obtain coverage in the extraction patterns of all of the ways in which the semantics are expressed by students. As the system is used, it logs all transactions and records student speech. When tutorials are deployed for live use, all session data are uploaded to a server each night. The data are processed automatically to assess system confidence in the interpretation of student responses. Using an Active Learning paradigm, low confidence sessions are selected for transcription and annotation. Once annotated, the data are added to the training set and system models (acoustic models, language models and extraction patterns) are retrained. Periodically, data are sampled for test sets and a learning curve is plotted for each module. All elements of this process are automatic except for transcription and annotation.

Generating Tutor Moves

The virtual tutor has a set of resources to conduct the session dialog; synthesized prompts, recorded prompts, narrations, static visuals, silent animations, narrated animations and interactive simulations. The tutor model controls how the resources for each tutor turn are selected. Features used for move selection include; semantic representation of the last prompt, whether the student reply was responsive to the prompt, and a comparison between the extracted representation from student responses and the reference representation from the narrative. These features generally express whether each target frame role a) hasn’t been addressed, b) has been prompted for but not answered, c) has been expressed incorrectly or d) has been correctly expressed. Boolean expressions of features are used to select the next tutor move. Tutor moves are sequences of the basic tutor actions: speak(play a recorded audio file), synthesize(a specified word string), flash(execute Flash application) and play(static media file or recorded video). Production rules in the form of Boolean expressions of features are associated with a sequence of actions to be taken by the tutor if the rule evaluates true. Some example pattern-action rules are:

```
# last student response indicated boredom
Response == "boredom"
          Action: "synth(So, I have to be entertaining every minute? You try it some time.)”

# Got it all right, give positive feedback and re-state
Origin == Reference:Origin AND Destination == Reference:Destination
          Action: "synth(Excellent observations!);
                    synth(So, electricity is flowing from the negative end of the battery
                              and back to the positive end of the battery)”

# origin wrong
Origin != Reference:Origin
          Action: "synth(Let's take a look at something together. Look at the flow of electricity.
                            What do you notice about which end the electricity is flowing away from?)”
```

Templates are created for interaction types to make authoring of dialog interactions more efficient. For example, when discussing word definitions, set membership, and causal relations, very similar dialog sequences are used regardless of specific content. This is especially true of the introductory parts of each concept, where very open-ended prompts are used. *Tell* types of moves introduce a concept and present a narrated animation. *Elicit* type moves might make an opening statement to segue into a concept, present a silent animation and ask “What’s going on here?”. Elicitation of explanation of a causal relationship might use a scenario using and interactive simulation. Ask “What do you think would happen if …”, then have the student try it in the simulation and then explain their observation.” The specific predicates and entities are different, but the interaction pattern is very similar.

During initial development and testing of dialogs, synthetic speech is used in the virtual tutor to allow easy modification. The application could use synthesis in field use, but we generally choose to have prompts recorded by a voice talent before students engage with Marni. This is a viable option since prompts for a session are known in advance and we have an efficient procedure for recording them. System tools generate the set of sentences to be recorded and a recording application is provided to efficiently manage recording and verifying each prompt, as well as the accuracy of the alignment of the speech to the movements of Marni’s lips and associating each audio file with the word string. The tools also automatically produce a task control file where all synth(word string) actions have been replaced with play(recorded file) actions.

Summary of Current MyST Tutorials Dialog Development Process

The primary activities involved in the development of MyST tutorial sessions are: development of Flash media, authoring feature expressions and associated action sequences and annotating data for extracting semantic representations. Templates of interaction types are used to reduce the effort of creating new tutor models. An efficient process is in place for collecting and annotating data and re-training system models.

Fifty tutorial sessions were developed in four months by a small team (one project manager, two digital artists and two linguistics students).

That optimistic assessment notwithstanding, substantial effort is required to develop and tune multi-media conversational tutorials. Less expensive media can be substituted for Flash animations, but the media is so integral to the presentation that we feel the expense justified. The other labor intensive effort is the annotation of extraction patterns. The next section details a proposal for reducing the data and effort required for training the semantic extraction model.

Applying Linguistic Resources to Semantic Extraction

One of the more costly and time-consuming aspects of developing a tutorial with this model is achieving good coverage in the extraction patterns used in parsing. The semantics of the domain are constrained, but student responses can vary greatly in the ways they choose to express concepts and terms. An efficient process is in place for collecting data and training the system, but the first time the system sees a construct it has not seen before, it does not extract it correctly. It still takes time, effort and data to get good coverage of student responses.

The patterns are used to extract (and normalize) entities into semantic roles, and thus represent both patterns for entity recognition and higher-level patterns assigning the entities to roles. Entity patterns represent the set of phrases considered to be an acceptable synonym for a term. Electricity could be expressed as *electricity, energy, power, current or electrical energy*. Coverage of term synonyms from annotated data is achieved fairly quickly and easily and can be done by most anyone familiar with the domain. The larger problem is the patterns discriminating between possible role assignments. Not only is there more disfluency and variability here, annotating them is a more difficult task for someone not trained to do it.

One possibility for increasing robustness of extraction patterns and reducing data (and effort) needed to achieve coverage for role assignment is to use output from a domain-independent semantic role labeling (SRL) system to help with role assignment. The Proposition Bank (PropBank) provides a corpus of sentences annotated with domain-independent semantic roles (Palmer, et.al.). PropBank has been widely used for the development of machine learning based Semantic Role Labeling (SRL) systems. Pradhan et. al. (2005) used the representation in open domain question answering and Albright et.al. (2013) extended PropBank for processing clinical narratives. The idea is not to try to use PropBank output directly to produce the extracted representations, but to map PropBank SRL output onto MyST frames Domain specific entity patterns will still need to be applied to produce the canonical extracted form, but this is a much simpler task than role assignment and one more suited to non-linguists.

An initial investigation has been conducted to examine how well the semantic frames used in MyST can be produced from PropBank roles. Many of the roles can be mapped directly, such as class membership. In some cases, such as causal relations between two events, several PropBank predicates are involved in producing the MyST frame. PropBank parses are oriented around a predicate and separate parses are produced for each predicate. These need to be unified to produce the MyST frame. An example of a Propbank parse that maps directly is:

All metals are conductors

PropBank	MyST
Predicate: are	Frame: ClassMembership
A1: metals	Member: metals
A2: conductors	Class: conductors

And an example of one that is not so direct is:

When the switch is closed electricity flows

PropBank	MyST
Predicate: flow	Frame: CausalRelation
A1: electricity	Cause: SwitchState: closed
TMP: when the switch is closed	Result: ElectricalFlow: on

The MyST patterns produce the *SwitchState: closed* and *ElectricalFlow: on* elements. The mapping issue is that Propbank treats *When the switch is closed* as a temporal expression while the MyST frame treats it as a pre-condition (the *Cause* role covers both cause and pre-condition concepts). As the number of frames in a MyST tutorial is small, generally less than 20, rule based mapping of Propbank predicates and roles to MyST frames seems feasible.

In MyST, many different related predicates share the same frame. Students could say electricity *flows, goes, runs, races, zooms*, or *circles*, and the important elements are *what* is moving, *from where*, *to where*, irrespective of the choice of verb. The goal is to map PropBank predicates that share similar role sets onto a common MyST frame to provide general ways of talking about the event participants e.g., a set of patterns for talking about roles in motion events. The following two sentences describe motion in two very different domains, but use the same semantic frame for representing the meanings:

Electricity is flowing from the negative terminal to the positive
Predicate: Move
Theme: Electricity
Source: from the negative terminal
Goal: to the positive

The clouds are blowing from the west to the east
Predicate: Move
Theme: clouds
Source: from the west
Goal: to the east

In MyST, the recognition and clustering of predicates is done by the extraction patterns. As an example, the predicate term *Move* might have synonyms, move, flow and circle around. This gives no guidance of what to do when a new predicate is encountered. For example, suppose a student says *Electrons are zipping around in a circle*, and the system has never encountered the word *zipping.* The extraction patterns do not indicate that *zipping* is a form of movement. A saving grace of the system is that a predicate is not required to extract into a frame. The system produces the set of possible extracted frames and uses context to disambiguate between competing alternatives. As long as the role assignments are not ambiguous (as in Source and Goal) it is often able to perform the semantic frame extraction correctly. Sometimes however, extraction patterns for roles do not cover the construction used by the student. Incorporating PropBank parses offers the possibility to save considerable annotation effort by doing role assignment in a domain-independent way so that extraction patterns are mostly only required to add structure to and normalize entities. It is expected that some MyST frames might not have a useful mapping from PropBank roles and will still require extraction patterns, but that most can be mapped from PropBank. At the current time, there is no quantitative data to support this, only a pilot investigation.

Adapting PropBank to Domain and Genre

Even though PropBank uses a domain-independent representation, machine learning based systems trained on it will necessarily be learning aspects of the topic and genre used in the training data. Initial PropBank training data were sentences taken from the Wall Street Journal and the Brown Corpus, both fluent written text. When PropBank trained SRL systems were applied to clinical narratives in the medical domain, both the genre of dictated notes and idiosyncratic word usage in the medical domain were very different from the original training data, and lowered performance (Albright et al., 2013). Parser performance was enhanced significantly by annotating a modest amount of data in the new domain with PropBank labels.

None of the available PropBank corpora are a good match to either topic or genre for children's conversational speech on science. There currently is no large corpus available that is appropriate for training PropBank parsers for spoken dialog based science tutorials for children. Boulder Language Technologies is beginning the work of annotating data collected in the MyST project to provide such a resource, representing over 1000 hours of speech from over 1200 elementary school students.

RECOMMENDATIONS AND FUTURE RESEARCH

While most of the mechanisms in the MyST framework are similar to capabilities that are already contained in GIFT, we believe that the extraction and use of domain specific sematic roles can provide complementary information to the current set of features being used. The functions for annotating data, training extraction patterns and extracting semantic frames could easily be integrated into the GIFT framework and the features derived from them made available as additional information within the current framework. The tools for selecting data for new annotations to add to the training data, and for evaluating component performance can be used to expand the representation as the systems evolve over time.

Boulder Language Technologies will make all of the components of the MyST system available for research use, including the Bavieca Automatic Speech Recognition engine, Phoenix Natural Language Processing engine and a character animation system. Many of these components are trained from data, and both supervised and unsupervised training can improve the models. Many projects have benefitted from the sharing of data within a research community. An example is the Linguistic Data Consortium, which serves as a repository and distribution center for corpora. The availability of corpora reduces the entry barrier to new research efforts to improve the technology. When corpora are available, common tasks can be defined and common evaluations conducted to accelerate progress in the field. The availability of data tends to attract new researchers. We recommend that providing methods for sharing data by GIFT users, including common annotation guidelines and assessment conventions, be considered.

REFERENCES

Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W., Warner, C., Hwang, J., Choi, J., Dligach, D., Nielsen, R., Martin, J., Ward, W., Palmer, M., Savova, G. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. *JAMIA*, 20(5), 922-930.

Baker, C., Fillmore, C., & Lowe, J. (1998). The Berkeley FrameNet project. In Proceedings of the COLING-ACL, 86-90.

Bakhtin, M. (1975). *The dialogic imagination*. Austin, TX, University of Texas Press.

Bakhtin, M. (1986). *Speech genres and other late essays*. Austin, Tx, University of Texas Press.

Beck, I., & McKeown, M. (2006). *Improving comprehension with Questioning the Author: A fresh and expanded view of a powerful approac.*, New York: Scholastic.

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

Bricker, L., & Bell, P. (2009). Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. *Science Education, 82*, 473-498.

Bricker, L. A., & Bell, P. (2014). What comes to mind when you think of science? The perfumery!: Documenting science-related cultural learning pathways across contexts and timescales. *Journal of Research in Science Teaching, 51*(3), 260-285. doi: 10.1002/tea.21134.

Chi, M.T.H. (2009) Active-contructive-interactive: a conceptual framework for differentiating learning activities. *Topics in Cognitive Science,* 1:73-105.

Chin, C., & Osborne, J. (2010). Students' questions and discursive interaction: Their impact on argumentation during collaborative group discussions in science. *Journal of Research in Science Teaching, 47*(7), 883-908. doi: 10.1002/tea.20385.

deJong, L., Zacharia (2013). Physical and virtual laboratories in science and engineering education. *Science, 340*(305).

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education, 32,* 268-291.

Duschl, R., Schweingruber, H., & Shouse, A. (2007). *Taking science to school: Learning and teaching science in grades K-8*: National Academy Press.

Erduran, S., & Aleixandre, M. (2008). *Argumentation in science education: perspectives from classroom-based research*: Springer.

Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22*(4), 39-51.

Kelly, G., Regev, J., & Prothero, W. (2008). Analysis of lines of reasoning in written argumentation. In S. Erduran & M. P. Jimenez-Aleixandre (Eds.), *Argumentation in science education: Perspectives from classroom-based research*. New York: Springer.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education, 77*(3), 319-337.

Kuhn, D. (2010). Teaching and learning science as argument. *Science Education*, 94, 810–824. doi:10.1002/sce.20395.

Kulatunga, & Lewis. (2013). Exploration of peer leader verbal behaviors as they intervene with small groups in college chemistry. *Chemistry Education Research and Practice, 14,* 576-588.

Kulatunga, U., Moog, R. S., & Lewis, J. E. (2013). Argumentation and participation patterns in general chemistry peer-led sessions. *Journal of Research in Science Teaching, 50*(10), 1207-1231. doi: 10.1002/tea.21107

Lehrer, R., Schauble, L., & Lucas, D. (1998). Supporting development of the epistemology of inquiry. *Cognitive development of mental representation - theories and applications, 23*, 512-529.

Lehrer, R., Schauble, L., & Petrosino, A. J. (2001). Reconsidering the role of experiment in science education. In K. Crowley, C. Schunn, & T. Okada (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 251-277). Mahwah, NJ: Erlbaum.

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

McNeill, K. L. (2011). Elementary students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year. *Journal of Research in Science Teaching, 48*(7), 793-823. doi: 10.1002/tea.20430

McNeill, K., Lizotte, D., Krajcik, J., & Marx, R. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences, 15*(2), 153-191.

Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: an overview of project LISTEN. In K. Forbus & P. Feltovich (Eds.), *Smart machines in education* (pp. 169-234). MIT Press.

NAEP (2009), National and state reports in science *The nations report card*: National assessment of educational progress from http://nces.ed.gov/nationsreportcard

Naylor, S., Keogh, B., & Downing, B. (2007). Argumentation and primary science. *Research in Science Education, 37*(17), 39.

Nussbaum, E., Sinatra, G., & Poliquin, A. (2008). Role of epistemic beliefs and scientific argumentation in science learning. *International Journal of Science Education, 30*, 1977-1999.

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

Palmer, M., Gildea, D., & and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1), 71-106.

Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J., & Jurafsky, D. (2005). Support vector learning for semantic argument classification. *Machine Learning*, 60(1), 11-39.

Roth, W.-M. (2013). An integrated theory of thinking and speaking that draws on Vygotsky and Bakhtin/Vološinov. *Dialogical Pedagogy, 1*, 32–53.

Roth, W.-M. (2014). Science language *Wanted Alive*: Through the dialectical/dialogical lens of Vygotsky and the Bakhtin circle. *Journal of Research in Science Teaching, 51*, 1049–1083. DOI: 10.1002/tea.21158

Sampson, V., & Clark, D. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education, 92*(3), 447-472.

Sampson, Grooms, J., & Walker, J. (2009). Argument-Driven Inquiry: A way to promote learning during laboratory activities. *The Science Teacher, 76*(7), 42-47.

Schworm, & Renkle. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology, 99*(2), 285-296.

Simon, S., Erduran, S., & Osborn, J. (2006). Learning to teach argumentation: Research and development in the science classroom. *International Journal of Science Education,* 235-260.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. *Educational Psychologist*, 46(4), 197-221.

Voss, J., & Means, M. (1991). Learning to reason via instruction in argumentation. *Learning and instruction, 1*(337-350).

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

Vygotsky, L.S. (1987) Thinking and Speech. In R.W. Rieber & A.S. Carton (Eds.) *The collected works of L.S. Vygotsky, Vol. 1, Problems of general psychology.* (N. Minick, Trans.) (pp.39-285) New York: Plenum Press.

Ward, W., Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., van Vuuren, S., Weston, T., & Zheng, J. (2011), My science tutor: A conversational multi-media virtual tutor for elementary school science. *ACM Transactions on Speech and Language Processing*, 7(4).

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., & Weston, T. (2013), My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology, 105*, 1115-1125.doi: 10.1037/a0031589

Wise, B., Cole, R., Van Vuuren, S., Schwartz, S., Snyder, L., Ngampatipatpong, N., Pellom, B. (2005). Learning to read with a virtual tutor: foundations to literacy. In C. Kinzer & L. Verhoeven (Eds.), *Interactive Literacy Education*, Lawrence Erlbaum, Mahwah,NJ

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching, 39*(1), 35-62. doi: 10.1002/tea.10008

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 50 (2018) 157–174

www.elsevier.com/locate/csl

One-on-one and small group conversations with an intelligent virtual science tutor[▮]

Ronald Cole[a,*,1], Cindy Buchenroth-Martin[a], Timothy Weston[a], Liam Devine[a], Jeannine Myatt[a], Brandon Helding[a], Sameer Pradhan[a], Margaret McKeown[c], Samantha Messier[b], Jennifer Borum[b], Wayne Ward[a]

[a] *Boulder Learning, Inc., United States*
[b] *Boulder Valley School District, United States*
[c] *University of Pittsburgh, Learning, Research and Development Center, United States*

Received 3 May 2016; received in revised form 23 December 2017; accepted 18 January 2018
Available online 2 February 2018

Abstract

In this study we investigated students' conversations with a virtual science tutor (Marni), either individually or in small groups. These constituted two treatment conditions. Students were presented with narrated multimedia science problems and explanations followed by question-answer dialogs with the virtual tutor. Students who received either one-on-one or small group tutoring received the same set of multimedia presentations and questions posed by the virtual tutor. Students in the small group condition discussed their answer before one student from that group responded to the tutor. We asked if students receiving tutoring using the virtual tutor in groups would demonstrate learning gains equivalent to those of students receiving one-on-one tutoring. We also asked if both groups would demonstrate greater learning gains from pretest to posttest than students in business-as-usual (control) classrooms who did not receive supplemental tutoring. One hundred eighty-three (183) students (in 13 classrooms at 4 schools) participated in the study. Of the 183 students, 114 were randomly assigned to tutoring in small groups using Marni; and 69 students received one-on-one tutoring with Marni. When compared with the control group, effect sizes for were d ¼ 0.48 for the group tutoring condition and d ¼ 0.51 for the one-on-one tutoring condition. A two-way ANOVA suggested a main effect for tutoring group, F ¼ 16.8, d*f* (41,171), $p < 0.001$. In general, students reported benefiting from listening to one another, and from the small group interactions, even though they sometimes disagreed with the answer reported by the small group. We conclude our findings with a vision for a next generation of virtual science tutors that can facilitate discourse and argumentation among students in small groups, leading students to build on each other's ideas to construct accurate science explanations.

Keywords: Virtual; Tutoring; Intelligent; Learning; Student; Multimedia

[▮] This paper has been recommended for acceptance by Roger Moore.

* Corresponding author.

E-mail address: rcole@boulderlearning.com (R. Cole).

[1] www.boulderlearning.com.

http://dx.doi.org/10.1016/j.csl.2018.01.002
0885-2308/ 2018 Elsevier Ltd. All rights reserved.

1. Introduction

In this article, we describe a project in which we developed and evaluated a program to help students gain proficiency in scientific discourse leading to improved science achievement. Helping students learn how to engage in scientific discourse and argumentation has become a major focus of science education in the U.S. An influential report, *Taking Science to School: Learning and Teaching Science in grades K–8* (National Research Council, 2007), used evidence on child development and learning to advocate for four strands of scientific proficiency for all students. Specifically: "Students who understand 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" (p. 2).

Our study was designed to (a) engage students in one-on-one conversations with a virtual science tutor, and (b) engage students in small groups in conversations stimulated by questions posed by a virtual tutor. In both groups, students were presented with a set of narrated multimedia science presentations, and engaged in conversations about the science. We hypothesized that students who received either one-on-one or small group tutoring sessions with a virtual tutor would demonstrate greater science learning gains than students who received similar classroom instruction, but did not engage in one-on-one or small group tutoring sessions.

The article is organized as follows. Section 1 provides an overview of, and scientific rationale for, the study. Section 2 describes the two experimental conditions used in the 3rd, 4th and 5th grade classrooms, then presents and briefly summarizes the results. Section 3 discusses these results; Section 4 identifies and describes the main conclusions from the investigation, and describes implications for future work.

1.1. Overview of the study

Students participated in this study in one of two conditions. Students in one condition engaged in a series of one-on-one tutoring sessions with a virtual science tutor, Marni. In these sessions, Marni asked each student questions about science presented in narrated multimedia presentations, including follow-on questions designed to stimulate students' reasoning about the science and construct accurate explanations. Students in a second condition interacted with Marni in small groups, of 3 students. Students in small groups received the same multimedia presentations individually, but were asked to discuss the questions with the other students in their group before the student in control of the microphone during the session provided Marni with a spoken response to her question. We asked students in each condition about their learning experiences, and compared learning gains of students in each condition to students in business-as-usual (control) classrooms. These control classrooms received similar classroom instruction to treatment students, but did not receive supplemental tutoring with Marni. Our main research questions were:

1. Would students in both one-on-one and group conditions achieve significant and equivalent learning gains relative to students in control classrooms who did not receive tutoring?
2. Would students in small groups engage in meaningful discourse and argumentation, and report that those discussions were beneficial?

1.2. Scientific foundations

1.2.1. Theory and research on scientific discourse and argumentation

Historically, research in argumentation and collaborative discourse has acknowledged the strong influences of the theories of Vygotsky (1978, 1987) and Bakhtin (1975, 1986), who argued that all learning occurs in and is shaped by the social, cultural, and linguistic contexts in which they occur. Roth (2013, 2014) provided an excellent integration of Vygotsky's and Bakhtin's theories and their relevance to research on collaborative discourse. He argued that, when considered in the context of these theories, "currently available analyses of science classroom talk do not appear to exhibit sufficient appreciation of the fact that words, statements, and language are living phenomena, that is, they inherently change in speaking" (Roth, 2014, in online Abstract). The seminal writings of Vygotsky and Bakhtin have had a profound influence on subsequent research in discourse and argumentation, including (Wells' 1997, 2000, 2008) research on dialogic inquiry. We embrace this emphasis as a main focus of the current study: to learn whether children in small groups will become comfortable using their words to express, support, defend, reflect on, and modify their ideas during scientific discourse.

The past 25 years have witnessed remarkable growth in research on discourse and argumentation in education. Kuhn (1993, 2000) argued that “a conception of science as argument has become to be widely advocated as a frame for science education” (p. 1). Support for argumentation has codified into both a reform movement and framework for science education. It is supported by growing evidence of substantial benefits of explicit instruction and practice on the quality of students’ argumentation and learning (Chin and Osborne, 2010; Harris et al., 2006; Kulatunga and Lewis, 2013; Kulatunga et al., 2013; McNeill, 2011; Nussbaum et al., 2008; Sampson et al., 2009; Schworm and Renkl, 2007; Simon et al., 2006; Voss and Means, 1991). Evidence from these studies indicated that argumentation can be improved by providing professional development to teachers or knowledgeable students (Berland and Reiser, 2009; Bricker and Bell, 2008, 2014; de Jong and Zacharia, 2013).

Metanalyses of programs designed to foster discourse and argumentation in US elementary and middle school classrooms identified several interventions that improved student achievement in language arts and literacy over the course of a school year (Murphy and Edwards, 2005; Murphy et al., 2009; Soter et al., 2008). One of these programs, Questioning the Author (Beck and McKeown, 2006), described below, motivated the dialog moves produced by the virtual tutor in the present study.

A classic study by Hake (1998) compared pretest vs. posttest performance of a diverse sample of over 6500 high school and college students on a standardized test of conceptual knowledge of physics in two conditions: *traditional classes* that involved lectures and little or no interaction among students, and *interactive classes* where teachers stopped their lectures to ask students to discuss questions. Students in interactive classes demonstrated 48% learning gains relative to 24% learning gains in classrooms where teachers lectured but did not ask questions.

Moreover, a synthesis of over 250 research studies by Black and Wiliam (2009) indicated that administering formative assessments to students, and providing teachers and students with feedback on their performance, produced effect sizes 0.40−0.70 on standardized tests, relative to students in classrooms who were not administered formative assessments. Therefore, students benefited from feedback on their understandings of the science they were learning, and teachers benefited from feedback that informed their instruction.

Finally, a meta-analysis by Chi (2009) indicated that students whose instruction involves *interactive tasks* that included collaborative discourse and argumentation learned more than students whose learning involved *constructive tasks*, (e.g., classroom investigations and written reports), or *active tasks* (e.g., classroom science investigations). Menekse et al., (2013) obtained strong evidence that interactive instruction led to higher scores on deep reasoning questions relative to constructive, active, or passive instructional methods.

In sum, over three decades of scientific research indicate the importance of integrating discourse and argumentation into classroom instruction to improve student motivation and learning. There is strong evidence that when teachers are trained to initiate and manage classroom conversations in which students share, compare, and modify their ideas, those students and their teachers become more engaged and excited about learning, and overall student achievement improves.

1.2.2. Discourse in US classrooms

Large-scale studies of discourse in U.S. classrooms indicate that extended conversations, in which students do most of the talking, are rare (Nystrand and Gamoran, 1991). Over a period of 2 years, trained observers paid 4 visits to 58 8th grade and 54 9th grade English classes in US parochial and public schools in rural, suburban and urban settings. The observers recorded the amount of time spent on different instructional activities, and recorded and coded over 23,000 teacher and student questions to form a set of variables contrasting monologic and dialogic dialogs. Questions were coded for *authenticity* (a question was authentic if the answer was not known in advance by the asker) and *uptake* (previous answers were incorporated into new questions). The study found that “in virtually all classes, the teacher asked nearly all the questions; few about literature were authentic, and equally few followed up on student responses” (Nystrand, 1997, p. 44). Specifically, on average, there was less than 50 s of discussion per 8th grade class period and 20 s per 9th grade class period. Approximately 60% of all classes had no discussions; the single classroom with the most discussion averaged 2 min of it. Interestingly, there was a significant positive correlation between the amount of discourse in individual classrooms and student achievement in language arts. These results were replicated in a second, large-scale study (Nystrand et al., 1997). Extended conversations about science are also rare in science classrooms; as Osborne (2010) noted, “argument and debate are common in science, yet they are virtually absent in science education" (online abstract). The lack of discourse in U.S. classrooms is especially puzzling given the compelling evidence that effective programs have been developed, as discussed above, in which teachers engage students in discourse leading to significant gains in student motivation and learning (Murphy et al., 2009; Soter et al., 2008).

1.2.3. Benefits of individualized instruction: human tutoring

Over three decades of research have indicated that learning is most effective when students receive individualized instruction, either one-on-one, or in small groups. Bloom (1984) summarized studies that reported 2 Sigma learning gains for students who received one-on-one or small group tutoring, relative to students who received regular classroom instruction. Evidence that tutoring works has been similarly obtained from dozens of well-designed research studies and meta-analyses (Cohen et al., 1982) and positive outcomes were obtained in large-scale evaluations of specific tutoring programs (Slavin and Madden, 1989; Topping and Whiteley, 1990).

Factors that contributed to the effectiveness of individualized instruction, measured by gains in student achievement were associated with tutoring aligned with classroom instruction, asking students authentic questions, scaffolding learning with hints after questions, follow-on questions, and media designed to stimulate reasoning and help students' build on prior knowledge. These factors helped students continually generate explanations in response to deep reasoning questions (Butcher, 2006; Chi et al., 1989; Craig et al., 2000; Driscoll et al., 2003; King, 1989; King et al., 1998; Palinscar and Brown, 1984; Pine and Messer, 2000; Soter et al., 2008). Hausmann and VanLehn (2007) noted that: "explaining has consistently been shown to be effective in producing robust learning gains in the laboratory and in the classroom" (p. 418).

1.2.4. Benefits of individualized instruction: intelligent tutoring systems

Research in intelligent tutoring systems addresses a current, critical need to provide teachers and students with accessible, inexpensive, and reliably effective tools for improving young learners' interest and achievement. They enhance learning by providing students with individualized and adaptive instruction like that provided by an effective human tutor. Most Intelligent Tutoring Systems (ITS) rely on typed input to support dialogs with high school and college students. Intelligent tutoring systems that support spoken dialogs with users include D'Mello, Graesser & King, 2010; Johnson, 2010; Johnson, Rickel & Lester, 2015; Litman & Pan, 2002, Littman & Silliman, and Ward et al., (2011; 2013).

Advances in research and development of Intelligent Tutoring Systems (ITS) have informed the design of systems that produce learning experiences and outcomes equivalent to human tutoring (e.g., Graesser, Chipman, Haynes & Olney, 2005; VanLehn, Lynch, Schulze, Shelby, Taylor, Treacy, Weinstein, & Winterhall, 2005). A recent review by Van Lehn (2011) compared learning gains in studies in which students received human tutoring or interacted with an ITS. Students in these studies participated in tasks that required problem solving and constructing explanations. When compared to students who did not receive tutoring, the effect size of human tutoring across studies was d¼0.79 whereas the effect size of ITSs was d = 0.76 (Cohen's d. Cohen, 1992). Van Lehn (2011) concluded that "intelligent tutoring systems are nearly as effective as human tutoring systems" (pg. 197). A meta-analysis of ITS by Ma, Adesope, Nesbit & Liu (2014) that incorporated 107 effect sizes involving 14,321 participants indicated that use of intelligent tutoring systems was associated with greater achievement, with moderate to large effect sizes, in comparison with teacher-led, large-group instruction (*g*¼ .42), non-ITS computer-based instruction (*g*¼.57) and textbooks or workbooks (*g*¼35; Hedges'*g*, Hedges, 1981). The analysis revealed there was no significant difference between learning from ITS and learning from individualized human tutoring.

The large majority of intelligent tutoring systems described in the scientific literature support typed input with high school students, college students, and adults. Intelligent tutoring systems that support spoken dialogs with users include D'Mello, Graesser & King, 2010; Johnson, 2010; Johnson, Rickel & Lester, 2015; Litman & Pan, 2002; Ward et al., (2011; 2013).

1.2.5. Benefits of peer collaboration in intelligent tutoring systems

The development of technologies and systems to assess and facilitate collaborative problem solving is a vital and growing area of research (Hao et al., 2015; Liu et al., 2015; Zapata-Rivera et al., 2016; von Davier et al., 2017). The evidence indicates that collaborative problem solving is typically superior to the performance of individuals working independently. For example, Hoa et al. (2015) compared the performance of 486 individuals and 278 teams (dyads) who collaborated on a volcano science task using a web-based simulation; performance of the teams was significantly higher than performance of individuals.

Recent studies of collaborative learning in intelligent tutoring systems have reported statistically equivalent learning gains when students engaged in one-on-one tutoring or worked with other students to solve problems. Studies have compared student learning in the two conditions for secondary or college students in physics (Hausmann et al., 2008); engineering design (Kumar et al., 2010); computer science (Harsley et al., 2016) mathematics (Diziol et al., 2010), and language learning (Toussas et al, 2014). Few studies have compared children's learning during one-on-one or peer tutoring using an intelligent tutoring system. Olsen et al. (2016) found equivalent learning gains by 4th and 5th grade students who used a tutoring system to learn fractions, either individually or in small groups. A subsequent study (Olsen et al., 2017) compared students who (a) received one-on-one tutoring, (b) participated in peer tutoring, or (c) received a combination of both one-on-one and peer tutoring. The group that participated in both one-on-one and collaborative tutoring sessions produced significantly greater learning gains than students in either of the one-on-one or collaborative tutoring conditions.

1.2.6. Benefits of spoken dialogs between children and a virtual science tutor

My Science Tutor (MyST) is an intelligent tutoring system that engages students in conversations with a virtual science tutor, leading to learning gains comparable to those obtained with expert human tutors. To our knowledge, MyST is the only system developed to date that supports spoken tutorial dialogs with children (aged 7–10). MyST dialogs are aligned with science concepts encountered in small-group, classroom science investigations, using the Full Option Science System (FOSS) program. FOSS is a kit- and inquiry-based program, used by over 1 million students in 100,000 classrooms in the U.S. (FOSS, 2007). A typical FOSS science module includes 4 major Investigations over an 8–10-week period. For example, the in the FOSS Magnetism and Electricity (M&E) module, the 4 Investigations are Magnetism, Serial Circuits, Parallel Circuits, and Electromagnetism. Within each Investigation, students in small groups (3–5 students) conduct 4 different *classroom science investigations*. Thus, in a typical FOSS science module, students participate in a total of 16 classroom science investigations. Each MyST tutorial dialog session was aligned to the vocabulary and concepts students encountered in a specific classroom science investigation. Soon after completing a classroom science investigation, consented students left their classroom, and engaged in a tutorial dialog session with the virtual tutor Marni about the science being taught in the classroom.

Questioning the author: A defining feature of tutorial dialogs in MyST was that Marni asked students open-ended questions about science presented with media. The media included static illustrations, silent animations, or interactive simulations. Marni asked questions, like "What's going on here?" or "What would happen if ?"). MyST's spoken dialog system analyzed students' spoken answers to the question to identify which points were expressed by the student. When students did not express each of the points required for a complete explanation, Marni asked follow-on questions, and optionally presented new media, to stimulate reasoning and scaffold learning.

Marni's "dialog moves" are based on Questioning the Author (QtA), an effective approach to classroom discourse used by hundreds of teachers in US classrooms. In a metanalysis of programs designed to foster classroom discourse, Murphy and Edwards (2005) and Murphy et al. (2009), QtA was identified as one of two approaches to classroom conversations, out of nine examined, that promote high-level thinking and comprehension of texts. QtA produced effect sizes of 0.63 on text comprehension measures, and 2.5 SD's on researcher-developed measures of critical thinking and reasoning.

A key point in QtA is that student-to-student interactions are valuable only if they involve students truly listening to each other and building on each other's ideas. QtA *dialog moves* are designed to do exactly this. The role of the teacher (or virtual tutor) is to model this process, so students listen to each other, reflect on what other students have said, and self-assess and revise their ideas. The teacher or tutor models the process of listening carefully, extracting meaning and making connections between students' ideas using established dialog moves which paraphrase or elaborate student's ideas. Questions are substantial and meaningful. The teacher does not ask questions like "How many centimeters are in a meter?" Instead they ask for connections among student contributions: "How does that fit in with what Wayne said?" Or, "So are you disagreeing with what Jennifer said?"

During development of My Science Tutor, we worked closely with Margaret McKeown, co-developer of QtA, to incorporate key principles of QtA into tutorial dialogs between Marni and individual students, and to train human tutors to use QtA dialog moves when tutoring students. The QtA dialog moves used most often in MyST tutorials were *marking* and *rejoicing* (Beck and McKeown, 2006). These two techniques required the system identify the student's dialog content (marking it) followed by repeating (revoicing) a paraphrase of the information back to the student as a part of the next question: "Cindy, you mentioned that electricity flows in a closed path. What else can you tell me about how electricity flows?"

Detailed descriptions of how knowledge is represented in MyST dialogs, how the system interprets students' answers to open-ended questions, and the selection of prompts that Marni produces in response to students' answers, are described in Ward et al. (2011, 2013) and Ward and Cole (2016). A 3-min video of a student conversing with Marni about the flow of electricity in a serial circuit can be found at NSF-STEMforAll (2016).

MyST summative evaluation results: A summative evaluation of MyST, in five different science topics, was conducted during the 2010–2011 school year in 3rd, 4th and 5th grades. Students were randomly assigned to one of three conditions: (a) business as usual classroom instruction using the FOSS program, (b) supplemental tutoring by human tutors (using QtA dialog moves), or (c) tutorial dialog sessions with Marni. Over the course of an 8-week science module, taught in fall, winter or spring trimesters, students engaged in 16 dialog sessions lasting about 20 min—about 5 h of spoken dialogs with Marni. In an average dialog session, both Marni and the student produced approximately 7 min of speech. Analysis of transcriptions of over 1000 dialog sessions revealed that the MyST spoken dialog system recognized approximately 90% of all correct answers produced by students. Results of the evaluation indicated that Students tutored by Marni demonstrated learning gains of about one-half of one grade, which were statistically equivalent to learning gains of students who received human tutoring, relative to students in control classrooms who did not receive tutoring as a supplement to classroom instruction.

1.3. Novel features of the proposed work

Our review of the literature indicates that the present study is the first to compare children's learning, individually and in small groups, during spoken conversations with a virtual science tutor. A key distinction between our previous studies using the MyST spoken dialog system (Ward et al., 2011; 2013), and the present study, is that students in our previous studies engaged in spoken dialogs with the virtual tutor *throughout an entire 15–20-min session*. The tutor asked open-ended questions about science presented in media, and students produced spoken answers to the questions. Each session concluded with Marni presenting a brief (30–60-s) summary of the key science concepts discussed.

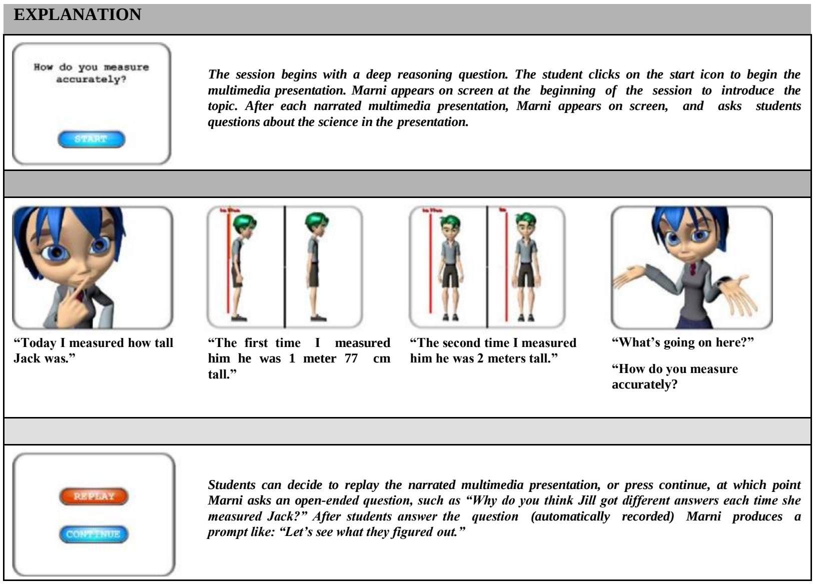
In the present study, individual students, or students in small groups, received narrated multimedia science presentations *before* engaging in spoken dialogs with the virtual tutor about the content of the presentations. In each session, students were presented with two narrated multimedia presentations. The first presentation introduced a science question or problem. Following the problem scenario, Marni asked students to explain the problem in their own words, with a question like "What was the problem Jack and Jill were trying to solve?" The second narrated presentation provided a solution to problem, followed by Marni asking students to explain the solution in their own words, e.g., So, how did Jack and Jill solve the problem?" Students in the small group condition were instructed to discuss their answers before the group leader (for the current session) presented a spoken answer to the tutor. Each»20-min session concluded with a 5–6-min spoken dialog with the virtual tutor. These dialogs were truncated versions of MyST spoken dialogs, with Marni asking questions about science presented in media about the key science concepts discussed in the session. Hereafter, we refer to spoken dialogs in our previous MyST studies (Ward et al., 2011, 2013) as MyST-SLS (Spoken Language System). We refer to the current study as MyST-MP&D (Multimedia Presentations and Dialogs).

2. The study

2.1. Sequence of MyST-MP&D activities

2.1.1. Title screen

Each MyST-MP&D session began with a title screen that presented a deep reasoning question (Fig. 1). In all cases, the printed question was read aloud by Marni. 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. This corresponds with research that has indicated that



EXPLANATION

The session begins with a deep reasoning question. The student clicks on the start icon to begin the multimedia presentation. Marni appears on screen at the beginning of the session to introduce the topic. After each narrated multimedia presentation, Marni appears on screen, and asks students questions about the science in the presentation.

"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?

Students can decide to replay the narrated multimedia presentation, or press continue, at which point Marni asks an open-ended question, such as "Why do you think Jill got different answers each time she measured Jack?" After students answer the question (automatically recorded) Marni produces a prompt like: "Let's see what they figured out."

Fig. 1. Deep reasoning question and problem scenario.

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 et al., 2010).

2.1.2. Engaging real-life scenario

The first multimedia presentation in each session was a narrated sequence of illustration that introduced the science problem (Fig. 1). It associated the science with materials and situations likely to be familiar to students, based on the content of the FOSS program, and our observations of classroom instruction. The Scenario was followed by an open-ended question by Marni, like "What's going there?" or "How are do you think Jack and Jill will solve the problem?" The real-life scenarios and Marni's questions were designed to help students make meaningful connections between the science content and their own experiences and knowledge, to introduce and discuss scientific vocabulary and concepts, and to help them make connections between the scenario they were provided, and the deep reasoning question introduced on the title screen.

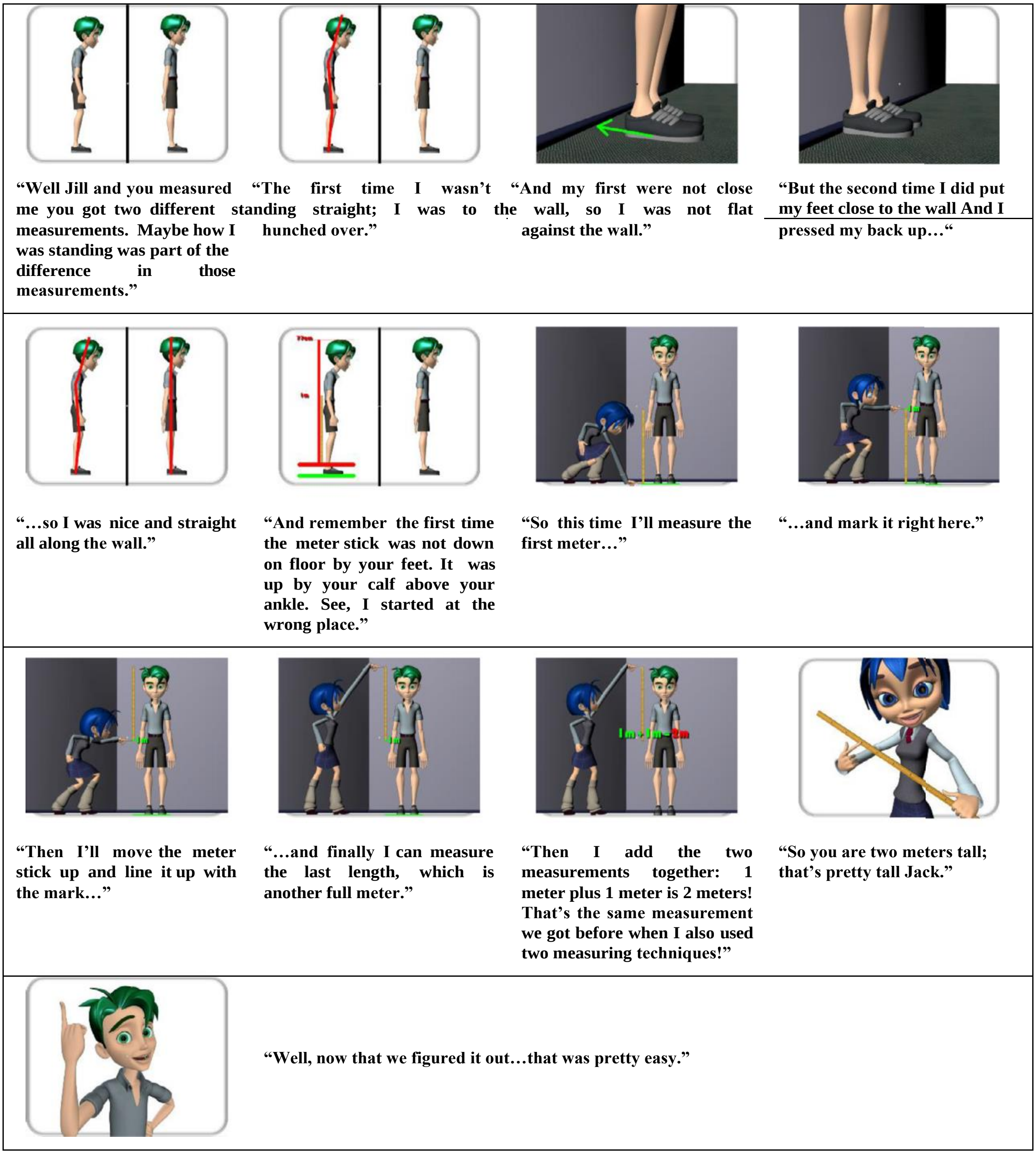
2.1.3. Multimedia science explanation

Students were presented with a multimedia presentation that explained the science content (Fig. 2). The design of the narrated multimedia presentations was informed by a substantial body of theory and research in multimedia learning. We sought to optimize learning and support the development of rich multimodal mental models that integrated verbal and visual information, leading to deep learning and transfer of knowledge to new contexts (Mayer, 2005). Presentations were sequenced, when necessary, with brief pauses between the key ideas so students had time to process each of key points in the presentations.

2.1.4. Formative assessment

After the multimedia presentation was completed, students were presented with an authentic question that could be answered if students had achieved a deep understanding of the science content. The question was sometimes the same as the deep reasoning question that introduced the overall MYST-MP&D session. In some tutoring sessions, however, a different question was presented. Questions were often accompanied by illustrations, and required answers that demonstrated application of science content knowledge to the situation shown in the picture with the newly provided, deep reasoning question.

2.1.4.1. Spoken response to the authentic question. Following presentation of the authentic about the science solution, students were asked to produce a spoken answer to the question. The goal was make students think about the



“Well Jill and 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…“

“…so I was nice and straight all along the wall.”

“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.”

“So this time I’ll measure the first meter…”

“…and mark it right here.”

“Then I’ll move the meter stick up and line it up with the mark…”

“…and finally I can measure the last length, which is another full meter.”

“Then I add the two measurements together: 1 meter plus 1 meter is 2 meters! That’s the same measurement we got before when I also used two measuring techniques!”

“So you are two meters tall; that’s pretty tall Jack.”

“Well, now that we figured it out…that was pretty easy.”

Fig. 2. Problem solution.

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 a single explanation.

2.1.4.2. Presentation of answer choices. After producing a spoken response to the question, students were presented with the question a second time, along with four possible answer choices. Students listened to Marni read each answer choice aloud before they could select the best answer. Answer choices were read aloud to assure that students considered all choices, since sometimes several answer choices were technically correct, but were not the best (e.g., most complete) answer to the question. After an answer was selected, Marni immediately provided formative feedback on the answer choice that was selected. If an incorrect answer was selected, Marni explained why it was incorrect, presented the correct answer, and expanded upon/explained why it was the correct one.

2.1.5. Spoken dialogs with Marni

Each session concluded with a spoken dialog with Marni lasting 4–6 min. These dialogs were condensed versions of MyST-SLS dialogs described in Ward et al. (2011, 2013), in which Marni asked open-ended questions about science presented on-screen as illustrations, silent animations, or interactive simulations. The questions, which were designed to stimulate reasoning and scaffold accurate explanations, addressed key points presented in the problem and solution scenarios, *in addition to* the key science concepts taught in classroom instruction about Measurement and Electricity and Magnetism. The Phoenix dialog system in MyST represented precise relationships among objects and events, such as electricity flows *out of* the negative pole of a D-cell and *into the* positive pole. Answering like “What do the poles of the D-cell have to do with the flow of electricity?” stimulated reasoning and discussion among students.

During tutorial dialogs, the Phoenix Spoken System in MyST) I in during tutorial dialogs with Marni.

2.2. Research questions

To quickly recapitulate the study, the two hypotheses 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 demonstrate greater learning gains than students without supplemental tutoring (based on pretest vs posttest scores).

We did not expect statistically significant different learning gains between group and one-on-one tutoring students; rather, we expected both groups to benefit from tutoring, and achieve gains similar to those obtained in previous studies.

2.3. Research design and data collection procedures

The assessment of the MyST-MP&D treatments was conducted from November of 2011 to May of 2012 in 3rd, 4th, and 5th grade classrooms in the Boulder Valley School District (BVSD). All students in the study received in-class instruction in either the FOSS module *Magnetism and Electricity (4th grade)* or *Measurement (3rd grade)*. Participating teachers followed module-based lesson plans and had their students conduct all science investigations as stipulated by the FOSS curriculum. The duration of instruction using the FOSS science modules varied from 1 to 3 months during the school year.

One hundred eighty-three (183) students in 13 classrooms at 4 schools participated in the study. Of the 183 students, 114 were randomly assigned to Group tutoring, and 69 were assigned to one-on-one tutoring with 100 students completing the FOSS *Magnetism and Electricity* module and 83 students completing the *Measurement* module. Students in the one-on-one condition interacted directly with Marni by answering questions verbally, or by selecting answers to multiple choice questions.

Students in the group tutoring condition worked in groups of 3 (except when a student was absent). Students were randomly assigned to groups across successive sessions. Each student in a group sat around a small table with a clear view of a single laptop computer. Students wore headphones, so they could look at and listen to Marni when she talked, and view and listen to the narrated, multimedia presentations. Students in groups were encouraged, at the beginning of each session, to discuss answers to Marni’s questions. When a discussion was concluded, the group leader pushed and held down the space bar on the laptop computer, and provided a spoken answer to Marni. Students took turns being the group leader across sessions. The MyST-MP&D system did not record individual student’s speech during discussions among students in small groups. Project tutors observed each group session, and coded students’ conversations, minute by minute, as discussed below.

2.3.1. Data and data cleaning

The FOSS assessing science knowledge (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. ASK assessments were developed and validated by FOSS and were keyed to the content in each module. Each module test has a different number of open-ended, short answer, multiple choice and graphing items. Learning gains from pretest to posttest for students in the individual and small group tutoring treatment conditions in MyST-MP&D 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. There were no significant demographic changes in free and reduced lunch percentages that occurred between years at participating schools. Also, no obvious changes happened in instruction on FOSS modules between years.

Because module tests had different scales, scores were standardized with 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 data) from each student’s score. This difference was then divided by the average standard deviation for both pre- and post-test score. Information about each test is presented in Table 1.

Pairs of raters scored all assessments from tutored students. The raters were project tutors from Boulder Language Technology who were blinded to subjects’ treatment conditions, and whether the assessments they scored were pre- or post-tests. Raters calibrated by working together on tests from the study using the scoring rubrics provided by FOSS, followed by independently scoring the assessments. Assessments jointly scored for assessment were re-graded by a different rater. Inter-rater reliabilities for two raters were high (counting only the open-ended items) with intra-class correlation coefficients ranging from 0.89 to 0.94, with averages for pre- and post-test being 0.91 and 0.94. Internal reliabilities (Cronbach’s Alpha) were lower, ranging from *a* ¼0.66 to *a* 0.87¼ for both pre- and post-test versions of the assessments, with averages for pre-test 0.78¼and post-test 0.78. ¼Internal reliability var- ied for each module. Scores used for outcome analyses were the averages across both raters.

2.4. Learning outcomes

The overall results are presented in Table 1, and Fig. 3 (right panel). Table 1 also includes the results from the MyST-SLS study conducted the previous year. When compared with the business-as-usual group, effect sizes were *d* ¼0.48 for the group tutoring condition and 0.51 for the one-on-one tutoring condition. Table 1 indicates that results of one-on-one and small group tutoring were similar to those found in the previous year in the MyST-SLS study, which compared learning gains following one-on-one tutoring with the virtual tutor (*d*¼0.51) and learning gains following human tutoring, (*d*¼0.65), relative to students in control classrooms who did not receive supplementation tutoring (Ward et al., 2011, 2013). In the current study, pre- versus post-test gain differences among

Table 1
Means, standard deviations (SD), sample sizes (*n*), and effect sizes for tutoring groups across the two years: 2011 and 2012.

Tutor condition	Mean	SD	*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	0.48
One-on-one MyST-MP&D (2012)	0.45	0.72	61	0.51

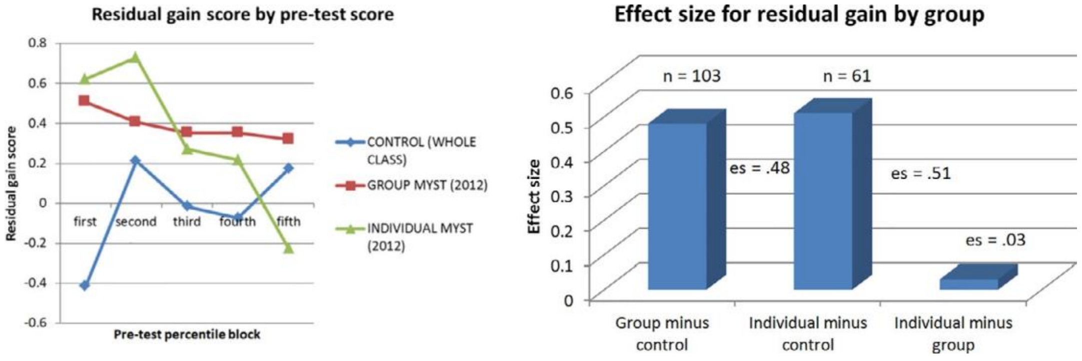


Fig. 3. Residual gain scores over time and experimental groups.

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.

We tested the same group differences with Kruskal–Wallis (KW) and Kolmogorov–Smirnov (KS) non-parametric tests using pre–post gain as a dependent variable. (Non-parametric tests have fewer statistical assumptions and usually provide a more conservative estimate of group differences.) The omnibus test for differences between all groups for both years was significant with KW¼44.9, $p < 0.0001$**. Results from the non-parametric comparisons were almost identical to the parametric tests, showing significant effects between tutoring conditions and the control for the first and second years, and no significant differences between MyST and Human tutoring conditions.

Learning gains were also assessed based on ability level on the pre-test, shown in Fig. 3 (left panel). Group comparisons divided the pre-test score distribution into 5 equal parts. The resulting distribution showed higher gains for tutored groups in the lower pretest score blocks especially for the one-on-one tutoring group, with more uniform gains across ability for students in the group tutoring condition (Fig. 3). Lower achieving students on the pre-test in the one-on-one tutoring group gained relatively more than students in the business-as-usual group, but gains for these students decreased for higher performing students on the pretest. Students in the group tutoring condition gained more than students in the business-as-usual group across the ability scale.

Comparisons with previous treatment groups from 2010 to 2011: A two-way parametric ANOVA tested if group means from both years differed significantly on residual gain scores. The main effect for tutoring group for all groups (2011, 2012) was statistically significant, *F*¼16.8, d*f*(41,171), $p < 0.001$. No statistically significant interaction was present for the treatment by module effect, indicating that differences found were generalizable to each module. Post-hoc tests showed statically significant differences between all tutoring groups and the business-as-usual group, and no significant differences among any of the four tutoring conditions. Effect sizes for MyST tutoring were higher in 2012 than 2011, although students that received face-to-face “human” tutoring demonstrated the largest gains, although differences between human and virtual tutoring conditions were not statistically significant.

2.5. Interactions among students in small groups

To review, the group consisted of one “leader” (the student who talked with Marni using a headset microphone) and 2 other students who discussed the answers with the leader but did not respond directly to Marni. 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. 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.

We observed how students in groups answered the virtual tutor’s 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 interactive 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 repeating it to Marni. Confirmational exchanges typically occurred following presentation of the answer choices to a multiple-choice question. Interactive discussions were usually longer in duration than confirmational exchanges, with students elaborating on each other's answers, disagreeing with each other, or referencing prior classroom instruction. A typical discussion had multiple back-and-forth exchanges culminating in an answer to which the students agreed and was spoken by the leader to Marni.

During structured observations of students as they interacted with MyST-MP&D, observers used a checklist to record the duration of student answers to questions from Marni, the types of questions asked by MyST, and characteristics of discussions between students. The checklist was completed using a portable digital assistant and electronic data were imported into an Excel database.

We tested the reliability of the observations by having two observers watch the same students on a minute-by-minute basis. 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 § 2 observations), and duration was highly correlated with observations and logs, ¼ 0.87. Data from two observers with low agreement and anomalous ratings were removed from the dataset. Five observers observed 64 students at 3 schools. Two hundred eight (208) tutoring sessions, 4749 group answers to questions, and 13,430 individual records were observed.[1]

2.5.1. Types of questions and responses during group discussions

Characteristics of interactive discussions: Interactive discussions were 54% of all types of conversations between students, and accounted for 65% of total time observed. These percentages varied widely across observations. The average interactive discussion was 30 s long (versus 16 s for the average confirmational exchange). When students did engage in interactive discussions, most 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 (revoicing). Fewer discussions involved students disagreeing with each other, occurring in 10% of discussions; students infrequently, in 3% of discussions, referred or referenced prior classroom instruction (this is an interesting result, given that the most of students reported that they sometimes disagreed with the answer the leader in each group ultimately provided to Marni on the follow-up questionnaire provided to each student).

We wanted to know if specific types of questions were more likely to elicit interactive discussions and confirmational exchanges. Students' responses were analyzed across 4 different types of questions:

1 *Initial questions*: These were the authentic questions that were asked immediately after presentation of the problem scenario, and the solution scenario.
2 *Multiple Choice (MC)* questions: Discussions students had about the 4 different answer choices that were read aloud to them, following the repeated spoken presentation of the authentic question that followed the solution scenario.
3 *Concept* questions during the final tutorial dialog session were designed to initiate assess student understanding of the key concepts to be discussed; e.g., "What can you tell me about how electricity flows through a serial circuit?"
4 *Follow-up* questions during the tutorial dialog session: If the group leader did not provide a complete and accurate answer to the concept question. Marni asked follow-up questions designed to elicit more precise answers about the specific science concepts; e.g., "What do the poles of the D-Cell have to do with the direction of flow of the electricity?"

Table 2 presents frequency counts for each type of interaction. The expected counts for each cell are the cross-tabular row and column products divided by the total number. Table 2 shows that the follow-up questions Marni asked during the 5-minute tutorial dialog that concluded each tutoring session accounted for 72% of all questions during the 20-minute tutoring session. On average, the group leader produced 27 spoken answers to Marni's questions

[1] Students were in groups of 2−4 students; records in the databases were organized by individual observations, observation sessions, and by students themselves.

Table 2
Frequency counts for question type by discussion and non-discussion interactions.

		Question type				Total
		Initial	MC	Follow- Up	Concept	
Non-discussion	Observed	505	454	4571	617	6147
	Expected	639	430	4425	653	6147
Discussion	Observed	867	470	4931	785	7053
	Expected	733 (+18%)	494 (¡ 5%)	5077 (¡ 3%)	749 (+6%)	7053
Total	Count	1372	924	9502	1402	13,200
	Expected Count	1372	924	9502	1402	13,200

during the tutorial dialog, with a standard deviation of 13.8. The average duration of each answer provided by the group leader was 3.0 seconds, or 1.5 minutes total speaking time during the 5-minute tutorial dialog. While follow-up questions were most frequent, they did not stimulate much group discussion relative to concept questions during the tutorial dialog, or the deep reasoning questions that followed the problem and solution scenarios.

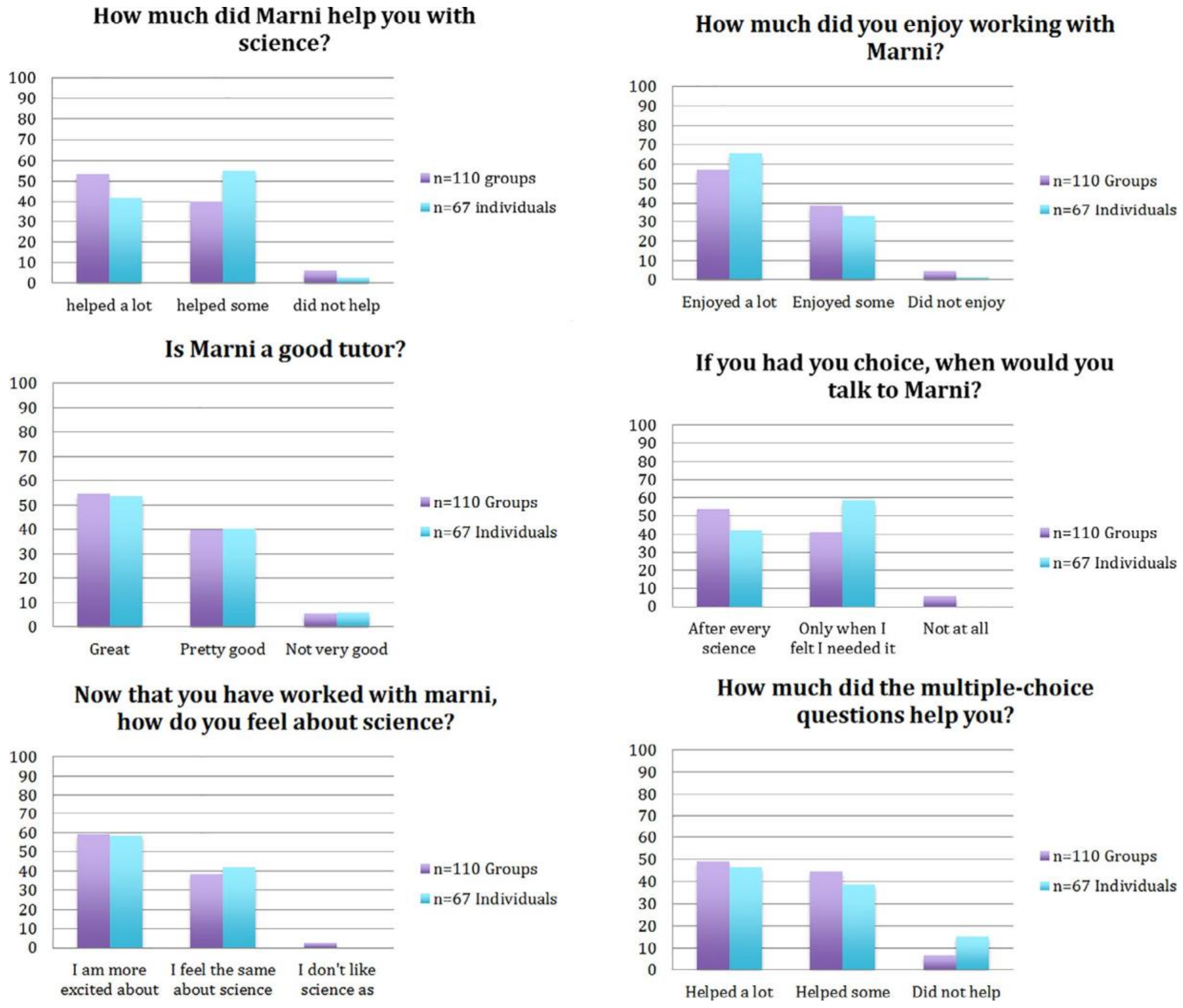


Fig. 4. Students' responses to survey questions in the small group and one-on-one tutoring conditions.

Our coded observations revealed that lengthier student discussions, with students elaborating on each other's answers, disagreeing about answers, or referencing classroom instruction, occurred more frequently when a) questions were asked immediately after the two narrated multimedia presentations, and b) questions were asked during the final spoken dialog after Marni asked the concept question. Extended discussions were less frequent for follow-on questions during the spoken dialogs, and for answer choices to multiple choice questions. The shortest group discussions were associated with students' answers to the alternative response choices to multiple choice questions. These were often confirmatory; the group quickly concurred with the answer choice selected by one of the members of the group. 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 achievement gains from pre- to post-test ASK assessment.

Links between FOSS-ASK assessments and types of responses in small groups: We also wanted to know if gains on the FOSS-ASK assessment were related to the frequency and duration of discussions. The average amount of time spent by students in discussions was correlated (*r*¼0.23) with pre-test scores, but not with either pre-test vs. post-test gains or post-test scores. This result generalized for both FOSS modules. The correlation with pre-test scores suggested that students who scored higher on the pre-test were more likely to engage in discussions.

2.6. Students experiences with MyST-MP&D

All students in both the one-on-one and group tutoring conditions in the MyST-MP&D study were administered the same written questionnaire. In addition, each student in the group tutoring condition responded to additional questions that were designed to gain insights about students' experiences when working with other students in those small groups. Fig. 4 indicates that students had very similar impressions in the two conditions.

Fig. 5 indicates that individual students in the group tutoring condition felt that they benefitted from their group discussions. And Interestingly, in response to the question: "How often did you disagree with the answer that the group gave to Marni?", over 60% said that the disagreed some of the indicated that they disagreed "sometimes." We believe this indicates that students were highly engaged in discussions, and maintained ideas and beliefs that were

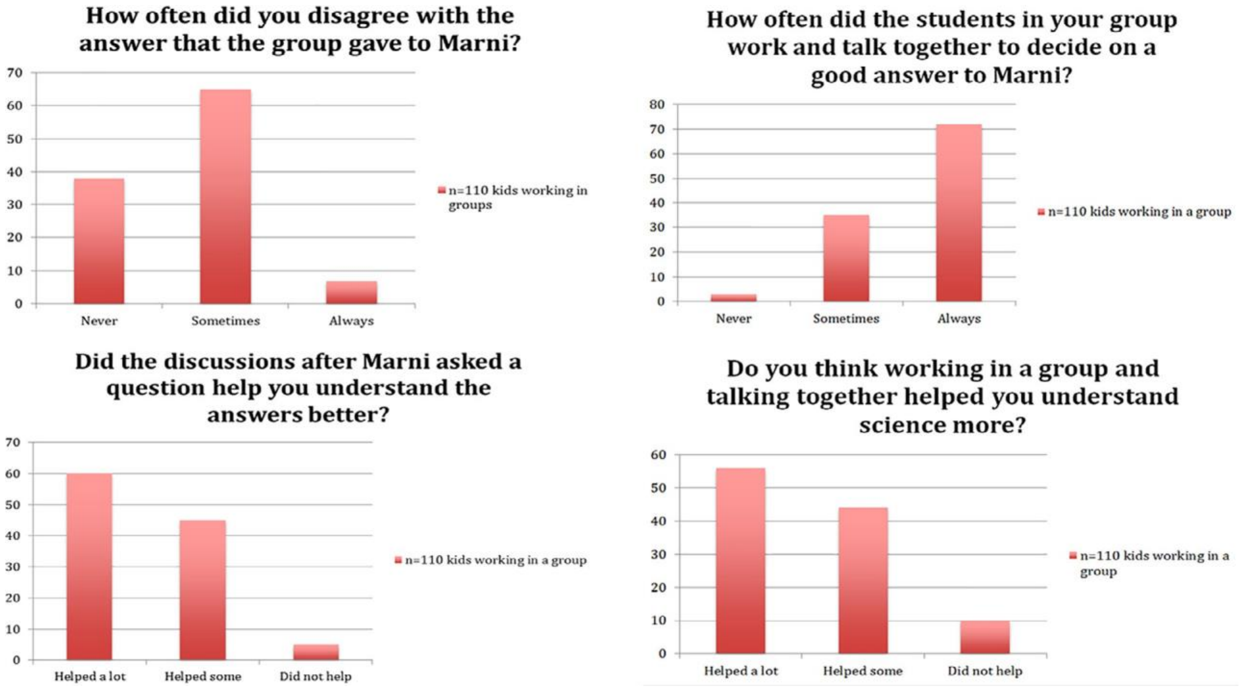


Fig. 5. Students' responses to surveys in small group tutoring condition.

not resolved to their satisfaction. Graesser and D'Mello (2012) argue that cognitive disequilibrium (and efforts to reduce cognitive dissonance) plays a central role in learning, including learning in intelligent tutoring systems. Students who reported they sometimes disagreed with the group answer may have experienced cognitive dissonance. We this interesting data point as evidence of students' strong engagement in group discussions.

3. Discussion

This study investigated student experiences and learning gains following one-on-tutoring sessions with a virtual science tutor, and student experiences and learning gains following small group tutoring sessions in which students were encouraged to discuss answers to questions posed by the virtual tutor before one of the students provided an answer for the whole group. The results indicated that learning gains were comparable for the 2 groups, and students in both groups had generally positive experiences. The magnitude of the learning gains was similar to those observed in previous MyST investigations (Ward et al., 2011, 2013).

In the current study, students in both one-on-one and group tutoring conditions were presented with a problem scenario, and then a solution to the problem. These sessions coherently presented science content with follow-up discussions, questioning, and explanations. Compared to MyST-SDS sessions, in which the session consisted mainly of listening to and answering questions posed by Marni, and viewing or interacting with media, the present study was more "self-contained" insofar as students were presented with problem scenarios and problem solutions, and asked questions about the problems and solutions that were presented within the tutoring session. Interestingly, similar learning gains were obtained in the MyST-SDS and MyST-MP&D studies. Further research is required to understand the relative contributions of incorporating narrated, multimedia problem scenarios and science explanations into tutorial dialog sessions, compared to tutoring sessions that consist mainly of asking questions about science encountered during classroom instruction. At this point, we have concluded that both approaches have and will produce comparable learning gains, when compared to students in business-as-usual classrooms (who receive similar classroom instruction, but do not receive supplementary tutoring). We also have concluded that the relationship between student achievement and particular aspects of the virtual tutoring any student received was complex and required additional investigation.

Also, the study produced the prime facie case for the plausibility of using virtual tutors to facilitate small group discussions. Almost without exception, students in small groups engaged in discourse and argumentation with their peers. In an initial pilot study, no prior instruction was provided to students on how to engage in effective discourse and argumentation or correctly provide scientific answers to complex problem scenarios. Rather, we took an ecological approach, in which we observed how students communicated with each other given the opportunity to share their responses and answers to questioning (scaffolding). Students reported that they enjoyed arguing with their peers, and that they believed they achieved a deeper understanding of the science because of the discussions. A main conclusion of the study is that 3rd, 4th and 5th grade students can and will engage in productive discourse, without explicit instruction about how to do so. Interestingly, 60% of student reported that they sometimes disagreed with the answer presented to the virtual tutor by the group leader following the group discussion.

In sum, this study demonstrated that students who received one-on-one tutoring with a virtual science tutor, and students who engaged in small group discussions following questions posed by the tutor, demonstrated the same significant gains in achievement, relative to students who did not engage in supplemental tutoring. This result was encouraging, as it motivates future research in which virtual tutors can be provided with sufficient intelligence to facilitate conversations during which students learn how to argue effectively, defend their arguments, and modify their ideas as they co-construct accurate science explanations with their peers.

4. Future work

We envision a near future in which a virtual tutor facilitates discussions among students in which they build on each other's ideas to construct explanations that lead to a deep understanding of the science. During conversations with the virtual tutor, speech recognition and natural language processing systems construct a semantic representation of each student's expressed understandings, gaps in their knowledge, and potential misconceptions. The tutoring system also builds a representation of the distributed knowledge of the entire group, based on correct answers produced by individual students. By creating a representation of individual student's expressed knowledge, and a unified representation of the group’s knowledge, the tutor can identify missing knowledge, the tutor can ask questions that help students listen to each other, consider each other’s arguments, modify their ideas, and work together to co-construct explanations that result in a deeper understanding of science for the entire group. We are confident that recent advances in Deep Learning algorithms will lead to graceful and effective dialogs between virtual tutors and groups of students that accurately model the conversational behaviors of highly effective teachers in physical or virtual classrooms.

Acknowledgment

This work was supported by funding from the National Science Foundation (Grant Number: 0733323), Division of Research on Learning in Formal and Informal Settings (DRL).

References

Bakhtin, M., 1975. The Dialogic Imagination. University of Texas, Austin, TX.

Bakhtin, M., 1986. Speech Genres and Other Late Essays. University of Texas Press., Austin, TX.

Beck, I., McKeown, M., 2006. Improving Comprehension with Questioning the Author: A fresh and Expanded View of a Powerful Approach. Scholastic, New York, NY.

Berland, L.K., Reiser, B.J., 2009. Making sense of argumentation and explanation. Sci. Educ. 93 (1), 26–55.

Black, P., Wiliam, D., 2009. Developing the theory of formative assessment. Educ. Assess. Eval. Account. (Formerly: J. Pers. Eval. Educ.) 21 (1), 5–31.

Bloom, B.S., 1984. The 2 sigma problem: the search for methods of group instruction as effective as one-on-one tutoring. Educ. Res. 13, 4–16.

Bricker, L.A., Bell, P., 2008. Conceptualizations of argumentation from science studies and the learning sciences and their implications for the practices of science education. Sci. Educ. 92 (3), 473–498.

Bricker, L.A., Bell, P., 2014. “What comes to mind when you think of science? The perfumery!”: documenting science-related cultural learning pathways across contexts and timescales. J. Res. Sci. Teach. 51 (3), 260–285.

Butcher, K.R., 2006. Learning from text with diagrams: Promoting mental model development and inference generation. J. Educ. Psychol. 98 (1), 182–197.

Chi, M.T.H., 2009. Active-constructive-interactive: a conceptual framework for differentiating learning activities. Top. Cognit. Sci. 1 (1), 73–105.

Chi, M.T.H., Bassok, M., Lewis, M.W., Reimann, P., Glaser, R., 1989. Self-explanations: how students study and use examples in learning to solve problems. Cognit. Sci. 13 (2), 145–182.

Chin, C., Osborne, J., 2010. Supporting argumentation through students’ questions: case studies in science classrooms. J. Learn. Sci. 19 (2), 230–284.

Cohen, J., 1992. A power primer. Psychological Bulletin 112 (2), 155–159.

Cohen, P.A., Kulik, J.A., Kulik, C.-L.C., 1982. Educational outcomes of tutoring: a meta-analysis of findings. Am. Educ. Res. J. 19 (2), 237–248.

Craig, S.D., Gholson, B., Ventura, M., Graesser, A., 2000. Overhearing dialogues and monologues in virtual tutoring sessions: effects on questioning and vicarious learning. Int. J. Artif. Intell. Educ. 11, 225–242.

de Jong, T., Linn, M.C., Zacharia, Z.C., 2013. Physical and virtual laboratories in science and engineering education. Science 340 (6130), 305–308.

D’Mello, S.K., Graesser, A., King, B., 2010. Towards Spoken Human-Computer Tutorial Dialogs. Human Computer Interaction. 25 (4), 289–323.

Diziol, D., Walker, E., Rummel, N., Koedinger, K., 2010. Using intelligent tutor technology to implement adaptive support for student collaboration. Educ. Psychol. Rev. 22 (1), 89–102.

Driscoll, D.M., Craig, S.D., Gholson, B., Ventura, M.Hu., Graesser, A.C., 2003. Vicarious learning: Effects of overhearing dialog and monolog-like discourse in a virtual tutoring session. J. Educ. Comput. Res. 29, 431–450.

FOSS. (2007). FOSSweb Login Page. Retrieved April 20, 2016, from http://www.fossweb.com.

Gholson, B., Witherspoon, A., Morgan, B., Brittingham, J.K., Coles, R., Graesser, A.C., Sullins, J., 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. Instr. Sci. 37, 487–493.

Graesser, A., Chipman, P., Haynes, B., Olney, A.M., 2005. AutoTutor: an intelligent tutoring system with mixed-initiative dialogue. IEEE Trans. Educ. 48 (4), 612–618.

Graesser, A.C., D’Mello, S., 2012. Emotions during the learning of difficult material. Psychol. Learn. Motiv. 57, 183–225 2012.

Hake, R.R., 1998. Interactive-engagement vs. traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am. J. Phys. 66, 64–74.

Hao, J., Liu, L., von Davier, A.A., Kyllonen, P., 2015. Assessing collaborative problem solving with simulation based tasks. In: Proceedings of the International Society of the Learning Sciences.

Harris, C.J., McNeill, K.L., Lizotte, D.J., Marx, R.W., Krajcik, J., 2006. Usable assessments for teaching science content and inquiry standards. Assessment in Science: Practical Experiences and Education Research, pp. 67–88.

Harsley, R., Di Eugenio, B., Green, N., Davide Fossati, D., Acharya, Sabita, 2016. In: Proceedings of 13th International Conference on Intelligent Tutoring Systems, ITS 2016 – Zagreb. Springer Verlag, Croatia.

Hausmann, R., van de Sande, B., VanLehn, K., et al., 2008. Shall we explain? Augmenting learning from intelligent tutoring systems and peer collaboration. In: Woolf, B. (Ed.), International Conference on Intelligent Tutoring Systems. Springer-Verlag, Berlin, Heidelberg, pp. 636–645.

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

Hedges, L.V., 1981. Distribution theory for Glass's, estimator of effect size and related estimators. Journal of Educational Statistics 6 (2), 107–128.

Johnson, W.L., Valente, A., 2009. A. Tactical Language and Culture Training Systems: Using AI to Teach Foreign Languages and Cultures.. AI Magazine 30 (2), 72–84.

Johnson, W.L., Rickel, J.W., Lester, J.C., 2000. Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments.. International Journal of Artificial Intelligence in Education 10 (1), 47–78.

King, A., 1989. Effects of self-questioning training on college students' comprehension of lectures. Contemp. Educ. Psychol. 14 (4), 366–381.

King, A., Staffieri, A., Adelgais, A., 1998. Mutual peer tutoring: effects of structuring tutorial interaction to scaffold peer learning. J. Educ. Psychol. 90 (1), 134.

Kuhn, D., 1993. Science as argument: implications for teaching and learning scientific thinking. Sci. Educ. 77 (3), 319–337.

Kuhn, D., 2000. Metacognitive development. Current Directions Psychol. Sci. 9 (5), 178–181.

Kulatunga, U., Lewis, J.E., 2013. Exploration of peer leader verbal behaviors as they intervene with small groups in college general chemistry. Chem. Educ. Res. Pract. 14 (4), 576–588.

Kulatunga, U., Moog, R.S., Lewis, J.E., 2013. Argumentation and participation patterns in general chemistry peer-led sessions. J. Res. Sci. Teach. 50 (10), 1207–1231.

Kumar, R., Ai, H., Beuth, J.L., Rosé, C.P., 2010. Socially capable conversational tutors can be effective in collaborative learning situations. In: Aleven, V., Kay, J., Mostow, J. (Eds.), Intelligent Tutoring Systems, ITS 2010. Lecture Notes in Computer Science. 6094, Springer, Berlin, Heidelberg.

Litman, D.J., Pan, S., 2002. Designing and evaluating an adaptive spoken dialogue system. User Modeling and User-Adapted Interaction 12, 111–137.

Litman, D. J., Silliman, S, (2004). ITSPOKE: An Intelligent Tutoring Spoken Dialog System. Proceedings of HLT-NAACL, Boston, Massachusetts, May 02 − 07, pp. 5−8.

Liu, L., Hao, J., von Davier, A.A., Kyllonen, P., Zapata-Rivera, D., 2015. A tough nut to crack: measuring collaborative problem solving. In: Rosen, Y., Ferrara, S. (Eds.), Handbook of Research on Technology Tools for Real-World Skill Development. Information Science Reference, Hershey, PA, USA, pp. 344–359.

Mayer, R. (Ed.), 2005. The Cambridge Handbook of Multimedia Learning. Cambridge University Press, Cambridge.

McNeill, K.L., 2011. Elementary students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year. J. Res. Sci. Teach. 48 (7), 793–823.

Menekse, M., Stump, G.S., Krause, S., Chi, M.T.H., 2013. Differentiated overt learning activities for effective instruction in engineering classrooms. J. Eng. Educ. 102 (3), 346–374.

Murphy, P.K., Edwards, M.N., 2005. What the Studies Tell Us: A Meta-Analysis of Discussion Approaches. In: Nystrand, M. (Ed.), American Educational Research Association.

Murphy, P.K., Wilkinson, I.A.G., Soter, A.O., Hennessey, M.N., Alexander, J.F., 2009. Examining the effects of classroom discussion on students' comprehension of text: a meta-analysis. J. Educ. Psychol. 101 (3), 740–764.

National Research Council, 2007. Taking Science to School: Learning and Teaching Science in Grades K-8. The National Academies Press, Washington D.C. (Committee on Science Learning Kindergarten through Eighth Grade). PDF available at: https://www.nap.edu/catalog/11625/taking-science-to-school-learning-and-teaching-science-in-grades.

NSF STEMforAll (2016). National Science Foundation Stem for All Video Showcase. https://vimeo.com/164599384.

Nussbaum, E.M., Sinatra, G.M., Poliquin, A., 2008. Role of epistemic beliefs and scientific argumentation in science learning. Int. J. Sci. Educ. 30 (15), 1977–1999.

Nystrand, M., Gamoran, A., Kachur, R., Prendergast, C., 1997. Opening Dialogue: Understanding the Dynamics of Language and Learning in the English Classroom. Teachers College Press, New York, NY.

Nystrand, M., 1997. Opening Dialogue: Understanding the Dynamics of Language and Learning in the English Classroom. Teachers College Press.

Nystrand, M., Gamoran, A., 1991. Instructional discourse, student engagement, and literature achievement. Res. Teach. Engl. 25 (3), 261–290.

Olsen, J.K., Rummel, N., Aleven, V., 2016. Investigating effects of embedding collaboration in an intelligent tutoring system for elementary school students. In: Looi, C.K., Polman, J., Cress, U., Reimann, P. (Eds.), Proceedings of the 12th International Conference of the Learning Sciences, ICLS 2016. I, International Society of the Learning Sciences, Singapore, pp. 338–345.

Olsen, J., Rummel, N., Aleven, V., 2017. Learning alone or together? A combination can be Best!. In: Proceedings of the 12th International Conference on Computer Supported Collaborative Learning, Philadelphia, PA.

Osborne, J., 2010. Arguing to learn in science: the role of collaborative, critical discourse. Science 328 (5977), 463–466.

Palinscar, A.S., Brown, A.L., 1984. Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. Cognit. Instr. 1 (2), 117–175.

Pine, K.J., Messer, D.J., 2000. The effect of explaining another's actions on children's implicit theories of balance. Cognit. Instr. 18 (1), 35–51.

Roth, W.-M., 2013. An integrated theory of thinking and speaking that draws on Vygotsky and Bakhtin/Vološinov. Dialogic Pedagogy: Int. Online J. 1, 32–53.

Roth, W.-M., 2014. Science language wanted alive: Through the dialectical/dialogical lens of Vygotsky and the Bakhtin circle. J. Res. Sci. Teach. 51 (8), 1049–1083.

Sampson, V., Grooms, J., Walker, J., 2009. Argument-driven inquiry. Sci. Teach. 76 (8), 42–47.

Schworm, S., Renkl, A., 2007. Learning argumentation skills through the use of prompts for self-explaining examples. J. Educ. Psychol. 99 (2), 285.

Simon, S., Erduran, S., Osborne, J., 2006. Learning to Teach Argumentation: Research and development in the science classroom. Int. J. Sci. Educ. 28 (2−3), 235–260.

Slavin, R.E., Madden, N.A., 1989. What works for students at risk: a research synthesis. Educ. Leadersh.: J. Dep. Superv. Curric. Dev., N.E.A 46 (5), 4–13.

Soter, A.O., Wilkinson, I.A., Murphy, P.K., Rudge, L., Reninger, K., Edwards, M., 2008. What the discourse tells us: talk and indicators of high-level comprehension. Int. J. Educ. Res. 47 (6), 372–391.

Sullins, J., Craig, S.D., Graesser, A.C., 2010. The influence of modality on deep-reasoning questions. Int. J. Learn. Technol. 5 (4), 378.

Topping, K., Whiteley, M., 1990. Participant evaluation of parent-tutored and peer-tutored projects in reading. Educ. Res. 32 (1), 14–32.

Troussas, C, Virvou, M., Alepsis, E., 2014. Collaborative learning: group interaction in an intelligent mobile-assisted multiple language learning system. Inform. Educ. 13 (2), 279–292.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J.A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., Wintersgill, M., 2005. The Andes physics tutoring system: Five years of evaluations. In: McCalla, G., Looi, C.K., Bredeweg, B., Breuker, J. (Eds.), Artificial Intelligence in Education. IOS Press, Amsterdam, Netherlands, pp. 678–685.

VanLehn, K., 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. 46 (4), 197–221.

Von Davier, A., Hao, J., Liu, L., Kyllonen, P., 2017. Interdisciplinary research agenda in support of assessment of collaborative problem solving: lessons learned from developing a collaborative science assessment prototype. Comput. Hum. Behav. 76, 631–640.

Voss, J.F., Means, M.L., 1991. Learning to reason via instruction in argumentation. Learn. Instr. 1 (4), 337–350.

Vygotsky, L., 1978. Mind in society: The development of higher psychological processes. Harvard University Press, Cambridge, MA.

Vygotsky, L., 1987. Thinking and speech. Plenum, New York, NY.

Ward, W., Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S.V., Becker, L., 2011. My science tutor: a conversational multimedia virtual tutor for elementary school science. ACM Trans. Speech Lang. Proces. 7 (4), 18.

Ward, W., Cole, R., Bolanos, D., Buchenroth-Martin, C., Svirsky, E., Weston, T., 2013. My science tutor: a conversational multimedia virtual tutor. J. Educ. Psychol. 105 (4), 1115–1125.

Ward, W., Cole, R., 2016. Developing conversational multimedia tutorial dialogs. In: Sottilare, R., Graesser, A., Hu, X., Brawne, K. (Eds.), Design Recommendations for Intelligent Tutoring Systems. Volume 3: Authoring Tools & Expert Modeling Techniques.. US Army Research Laboratory, pp. 243–254. Chapter 20.

Wells, G., 2000. Dialogic inquiry in education: building on the legacy of Vygotsky. In: Lee, C., Smagorinsky, P. (Eds.), Vygotskian Perspectives on Literacy Research. Cambridge University Press, New York, NY, pp. 51–85.

Wells, G., 1997. A sociocultural perspective on classroom discourse: Appendix A. Coding scheme for the analysis of classroom discourse. In: Davies, B., Corson, D. (Eds.), The Encyclopedia of Language and Education, 3, pp. 3–5.

Wells, G., 2008. Learning to use scientific concepts. Cultural Stud. Sci. Educ. 3 (2), 329–350.

Zapata-Rivera, D., Liu, L., Chen, L., Hao, J., von Davier, A., 2016. Assessing science inquiry skills in immersive, conversation-based systems. In: Daniel, B.K. (Ed.), Big Data and Learning Analytics in Higher Education. Springer International Publishing, pp. 237–252.

My Science Tutor: Learning Science with a Conversational Virtual Tutor

Sameer Pradhan **Ron Cole** **Wayne Ward**
Boulder Learning, Inc.
Boulder, CO
{pradhan,rcole,wward}@boulderlearning.com

Abstract

This paper presents a conversational, multimedia, virtual science tutor for elementary school students. It is built using state of the art speech recognition and spoken language understanding technology. This virtual science tutor is unique in that it elicits self-explanations from students for various science phenomena by engaging them in spoken dialogs and guided by illustrations, animations and interactive simulations. There is a lot of evidence that self-explanation works well as a tutorial paradigm, Summative evaluations indicate that students are highly engaged in the tutoring sessions, and achieve learning outcomes equivalent to expert human tutors. Tutorials are developed through a process of recording and annotating data from sessions with students, and then updating tutor models. It enthusiastically supported by students and teachers. Teachers report that it is feasible to integrate into their curriculum.

1 Introduction

According to the 2009 National Assessment of Educational Progress (NAEP, 2009), only 34 percent of fourth-graders, 30 percent of eighth-graders, and 21 percent of twelfth-graders tested as proficient in science. Thus, over two thirds of U.S. students are not proficient in science. The vast majority of these students are in low-performing schools that include a high percentage of disadvantaged students from families with low socioeconomic status, which often include English learners with low English language proficiency. Analysis of the NAEP scores in reading, math and science over the past twenty years indicate that this situation is getting worse. For example, the gap between English learners and English-only students, which is over one standard deviation lower for English learners, has increased rather than decreased over the past 20 years. Moreover, science instruction is often underemphasized in U.S. schools, with reading and math being stressed.

The Program for International Student Assessment (PISA), coordinated by the Organization for Economic Cooperation and Development (OECD), is administered every three years in 65 countries across the world. According to their findings in 2012, the U.S. average science score was not measurably different from the OECD average.

Our approach to address this problem is a conversational multimedia virtual tutor for elementary school science. The operating principles for the tutor are grounded on research from education and cognitive science where it has been shown that eliciting self-explanations plays an important role (Chi et al., 1989; Chi et al., 1994; Chi et al., 2001; Hausmann and VanLehn, 2007a; Hausmann and VanLehn, 2007b). Speech, language and character animation technologies play a central role because the focus of the system is on engagement and spoken explanations by students during spoken dialogs with the virtual tutor. Summative evaluations indicate that students are highly engaged in the tutoring sessions, and achieve learning outcomes equivalent to expert human tutors (Ward et al., 2011; Ward et al., 2013). Surveys of participating teachers indicate that it is feasible to incorporate the intervention into their curriculum. Also, importantly, most student surveys indicate enthusiastic support for the system.

Tutorials are developed through an iterative process of recording, annotating and analyzing logs from sessions with students, and then updating tutor models. This approach has been used to de-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations, pages 121–126,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

velop over 100 tutorial dialog sessions, of about 15 minutes each, in 8 areas of elementary school science.

My Science Tutor (MyST) provides a supplement to normal classroom science instruction that immerses students in a multimedia environment with a virtual science tutor that models an engaging and effective human tutor. The focus of the program is to improve each student's engagement, motivation and learning by helping them learn to visualize, reason about and explain science during conversations with the virtual tutor. The learning principles embedded in MyST are consistent with conclusions and recommendations of the National Research Council Report, "Taking Science to School: Learning and Teaching Science in Grades K-8" (NRC, 2007), which emphasizes the critical importance of scientific discourse in K-12 science education. The report identifies the following 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.

The report also emphasizes 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.

2 The MyST Application

MyST provides students with the scaffolding, modeling and practice they need to learn to reason and talk about science. Students learn science through natural spoken dialogs with the virtual tutor Marni, a 3-D computer character. Marni asks students open-ended questions related to illustrations, silent animations or interactive simulations displayed on the computer screen.

Figure 1 shows the student's screen with Marni asking questions about media displayed in a tutorial. The student's computer shows a full screen window that contains Marni, a display area for presenting media and a display button that indicates the listening status of the system. Marni produces accurate visual speech, with head and face movements that are synchronized with her speech. The media facilitate dialogs with Marni by helping students visualize the science they are discussing. The primary focus of dialogs is to elicit explanations from students. MyST compares the student's spoken explanations to reference explanations for the lesson by matching the extracted *semantic roles* using the Phoenix parser (Ward, 1991), then presents follow-on questions and media, to help the student construct a correct explanation of the phenomena being studied. The virtual tutor Marni, who speaks with a recorded human voice, is designed to model an effective human tutor that the student can relate to and work with to learn science. MyST provides a non-threatening and supportive environment for students to express their ideas. The dialogs scaffold learning by providing students with support when needed until they can apply new skills and knowledge independently.

MyST is intended to be used as an intervention for struggling students, with intended users being K-12 science students. While it should prove a benefit to all students, struggling students should benefit most. Depending on the recording conditions and ambient noise, as well as the characteristics of the student and session, the recognition word error rate ranges from low 20s to mid-40s. MyST will contain tutorials for 3 topics per grade, with content aligned with NGSS. For each topic, students engage in an average of 10 spoken dialog sessions with the tutor, lasting approximately 20 minutes each. oThe MyST tutorial sessions are in addition to the normal classroom instruction for the module. Tutoring sessions can be assigned as homework or during regular school hours, at the teacher's discretion. In the initial studies, tutoring was always done during regular school hours. Teachers specify the space in the school to be used, generally any relatively quiet room. Students are sent to use the system a few at a time, depending on how many computers are available (5 computers per classroom were used in the efficacy study). All students are given a demo at the beginning of the school year and given a chance to ask questions. Teachers schedule time for students, but students log on and use the system without supervi-

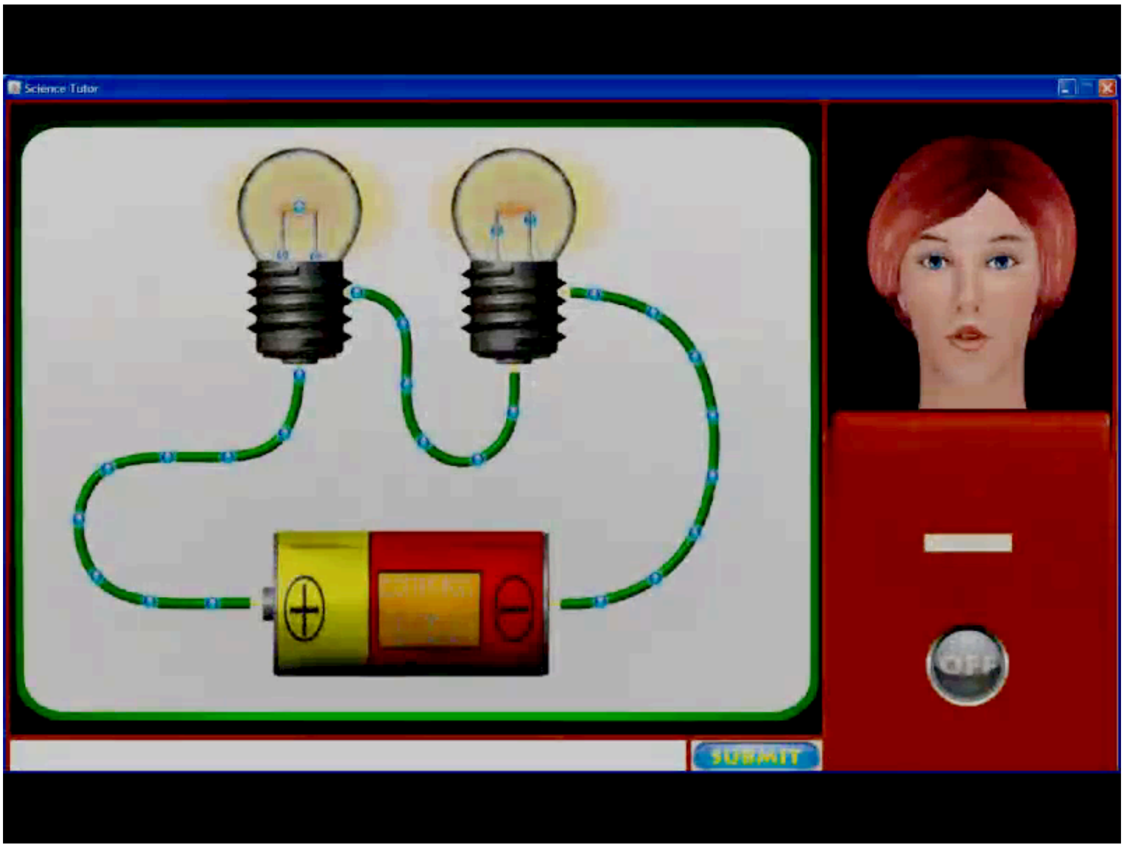


Figure 1: A snapshot of the screen as seen by a student.

sion, so it has minimal impact on teacher time or other human resources. In studies thus far, surveys report that teachers did not have problems using the system and it did not interfere with their other activities.

The application will eventually be deployed using a Software as a Service (SaaS) model. It will run on a server and students will access it through their browser. If internet service is not available or reliable, it can be run stand-alone and the data uploaded when service is available. Both content and user populations will evolve and system models need to incorporate dynamic adaptation in an efficient way. Data from all user sessions is logged in a database and is available for continuous evaluation and re-training of system models. The system is designed to work well even if it doesn't understand the user, but becomes more engaging and efficient as it understands the user better. As training data grows model parameters become more accurate and more explicit models are trained, such as acoustic models for ELL students. Unsupervised training is combined with active learning to optimize use of the data for tuning system models. Teachers in the initial studies did not feel that they would have a problem implementing the system.

3 Theoretical Framework

The theory of change, and theoretical and empirical support Science curricula are structured with new concepts building on those already encountered. Struggling students fall further and further behind if they don't understand the content of each topic. Research has demonstrated that human tutors are effective (Bloom, 1984; Madden and Slavin, 1989), media presentations are effective (Mayer, 2001) and QtA dialog strategies are effective (Murphy and Edwards, 2005). A system that emulates a human tutor using media presentations to focus a student's attention and conducting a QtA-style dialog with the student should also be effective. This additional time spent thinking and talking about the science concepts covered in class will enable students who would have fallen behind to understand the content of the current investigation so they will be prepared to participate in and understand subsequent topics. Student learning will increase because they are excited about and engaged by interesting and informative presentations that help them visualize and understand the science and because they will learn to engage in conversations in which they construct, reflect on and revise mental models and explanations about the science they are seeing and trying to explain. MyST dialogs are designed to provide students with understandable multimedia scenarios, explanations and challenges and a supportive social context for communication and learning. Science is introduced through scenarios that students can relate to and make sense of, and provide a context for introducing and using science vocabulary and making connections between vocabulary, objects, concepts and their prior knowledge. Multimedia learning tools show and explain science, and then enable students to revisit the media and explain the science in their own words.

Research has demonstrated that having students produce explanations improves learning (Chi et al., 1989; Chi et al., 2001; King, 1994; King et al., 1988; Palincsar and Brown, 1984). In a series of studies, Chi et al. (1989; 2001) 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 et al., 1994). Hausmann and Van Lehn (2007a; 2007b) note that: “self-explaining has consistently been shown to be effective in producing robust learning gains in the laboratory and in the classroom.” Experiments by Hausmann and Van Lehn (Hausmann and VanLehn, 2007a) indicate that it is the process of actively producing explanations, rather than the accuracy of the explanations, that makes the biggest contribution to learning.

4 Semantic Underpinnings

The patterns used in MyST to extract frames from student responses are trained from annotated data. The specification of tutorial semantics begins with creating a narrative. A tutorial narrative is a set of natural language statements that express the concepts to be discussed in as simple a form as possible. These do not represent the questions that the system asks, but are the set of points that the student should express.

The narrative represents what an ideal explanation from a student would look like. The narrative statements are manually annotated to reflect the desired semantic parses. These parsed statements define the domain of the tutorial. The initial grammar patterns are extracted from the narratives and have all of the roles and entities that will be discussed, but only a few (or one) ways of expressing them. As the system is used, the grammar is expanded to cover the various ways students articulate their understandings of the science concepts. This is done by annotating recordings of student responses generated in real use. So the life cycle of the natural language processing model for a module is:

1. Create and annotate a narrative to define the domain of the tutorial
2. Field the system to collect data from real users
3. Sample incoming data and annotate
4. Evaluate current model and re-train
5. Repeat step 3-4 as long as the module is used

As the system is used, it logs all transactions and records student speech. When tutorials are deployed for live use, incoming data are processed automatically to assess system confidence in the interpretation of student responses. High-confidence items are added to the training database, and low confidence sessions are selected for transcription and annotation. The system also provides a text input mode that students can use to interact with the Avatar. Once annotated, the data are added to the training set and system models (acoustic models, language models and extraction patterns) are retrained. Periodically, data are sampled for test sets and a learning curve is plotted for each module. All elements of this process are automatic except for transcription and annotation.

The semantics of each domain are constrained, but student responses can vary greatly in the ways they choose to express concepts and terms. It takes time, effort and data to get good coverage of student responses. Semantic annotation for the system consists of annotating:

Entities—*The basic concepts talked about in the session and the phrases that would be considered synonyms*. Electricity could be expressed as electricity, energy, power, current or electrical energy. Coverage of term synonyms from annotated data is generally achieved fairly quickly. **Roles**—*How the entities in an event or concept are related* *to each other*. The larger problem is to attain coverage of the patterns discriminating between possible role assignments. Not only is there more disfluency and variability here, annotating them is a more difficult task for someone not trained to do it. Currently, it takes about one hour for a highly-trained annotator to mark up the data collected in a single 20-minute tutorial session.

5 Extrinsic Evaluation

An assessment was conducted in schools to compare learning gains from human tutoring and MyST tutoring to business-as-usual classrooms. Learning gain was measured using standardized assessments given to students in each condition before and after each science module. Both tutoring conditions had significantly higher learning gains than the control group. While the effect size for human tutors vs. control (d=0.68) was larger than for MyST vs. control (d=0.53), statistical tests supported the hypothesis of no significant difference between the two.

A simple two-group comparison using a Repeated Measures ANOVA shows a statistically significant effect at F=46.4, df 1,759, p <.0001 favoring the treatment group. The interaction between group and module was also significant at F=9.5, p < .001. We also used an Analysis of Covariance (ANCOVA) to compare post-test scores. This procedure adjusts for pre-test differences while comparing the post-test average scores. The two-group comparison was significant at F=7.4, df 1,768, p=.018. We also saw a significant interaction between treatment group and module with F=12.4, df 3,768. Testing the main effects with a hierarchical mixed model with students nested within classrooms we found a significant effect for the treatment group at F=6.2, df 1,2l7,662, p=0.013. No significant interaction effect was found for module by group.

A written survey was given to the students who participated in the gas. Measures were taken to avoid bias wherein students give overly positive answers to questionnaires. The survey included questions that asked for ratings of student experience and impressions of the program and its usability. 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. Teachers were asked for anonymous feedback to help assess the feasibility of an intervention using the system and their perceptions of the impact of the system. A teacher survey was given to all participating teachers directly 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. Teachers answered items related to potential barriers in implementing new technology in the classroom. 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% indicated that they would like to have all of their students use the system (not just struggling students). Following these studies, Boulder Learning combined the best elements of the initial systems into the current MyST system, and with continued funding from IES (Cognition and Student Learning Goal 3), is conducting an efficacy study. We are currently in the 3rd year of a 4 year study. While data collection will continue for another year, preliminary results support the learning gain performance from the initial studies.

6 MyST Conversations Corpus of Student Speech (MCCSC)

We are making a cleaned up version of the corpus available to the research community[1] for free and for commercial use at a pre-determined cost. The first release of the corpus v0.1.0 comprises 298 hours of speech out of which 198 hours are manually transcribed. This covers roughly 1.4 million words of text. We are in the process of cleaning up about the same amount of collected data for future distribution.

7 Future Work

In the near future we plan to evaluate applying a statistical labeler trained on existing corpora to the task of Role assignment. This approach should provide increased robustness to novel input and substantially reduce the human annotation effort required to attain a given level of coverage. The

[1] `http://corpora.boulderlearning.com/myst`

Proposition Bank (PropBank) provides a corpus of sentences annotated with domain-independent semantic roles (Palmer et al., 2005). PropBank has been widely used for the development of machine learning based Semantic Role Labeling (SRL) systems. Pradhan et al. (2005) used a rich set of syntactic and semantic features to obtain a performance with F-score in the low-80s. It has been an integral component of most question answering systems for the past decade. Since its first application to the newswire text, PropBank has been extended to cover many more predicates and diverse genres in the DARPA OntoNotes project (Weischedel et al., 2011; Pradhan et al., 2013) and the DARPA BOLT program. We plan to map PropBank SRL output onto MyST frames. Domain specific entity patterns will still need to be applied to produce the canonical extracted form, but that is a much simpler task than role assignment and one more suited to non-linguists.

References

B. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16.

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

M. Chi, N. De Leeuw, M. Chiu, and C. LaVancher. 1994. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477.

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

R. G. M. Hausmann and K. VanLehn. 2007a. Explaining self-explaining: A contrast between content and generation. *Artificial Intelligence in Education*, pages 417–424.

R. G. M. Hausmann and K. VanLehn. 2007b. Self-explaining in the classroom: Learning curve evidence. In *29th Annual Conference of the Cognitive Science Society*, Mahwah, NJ.

A. King, A. Staffieri, and A. Adelgais. 1988. Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology*, 90(1):134–152.

A. 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).

N. A. Madden and R. E. Slavin. 1989. Effective programs for students at risk. In R. E. Slavin, N. L. Karweit, and N. A. Madden, editors, *Effective pull-out programs for students at risk*. Allyn and Bacon.

R. Mayer. 2001. *Multimedia Learning*. Cambridge University Press., Cambridge, U.K.

P. K. Murphy and M. N.b Edwards. 2005. What the studies tell us: A meta-analysis of discussion approaches. In *American Educational Research Association*, Montreal, Canada.

National Research Council. NRC. 2007. Taking science to school: Learning and teaching science in grades k-8. In R. A. Duschl, H. A. Schweingruber, and A. W. Shouse, editors, *Committee on Science Learning Kindergarten through Eighth Grade. Washington D.C.* The National Academies Press.

A. Palincsar and A. Brown. 1984. Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1(2).

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James Martin, and Dan Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria, August.

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

Wayne Ward, Ron Cole, Daniel Bolanos, C. Buchenroth-Martin, E. Svirsky, and Tim Weston. 2013. My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology*, 105(4):1115–1125.

W Ward. 1991. Understanding spontaneous speech: the phoenix system. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 365–367 vol.1, April.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.

C.2

NSF MyST Final Report

COLLABORATIVE RESEARCH:

IMPROVING SCIENCE LEARNING IN INQUIRY-BASED PROGRAMS

(NSF DRK-12 Grant 0733323)

Wayne Ward & Ron Cole

Boulder Language Technologies

Contents

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 Jyvaskyla, 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-Kołaczyk, 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 verses 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 Jyvaskyla 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 and 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 2010-2011assessment. 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 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.

Figure 1 – Virtual Tutor Screen

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

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?

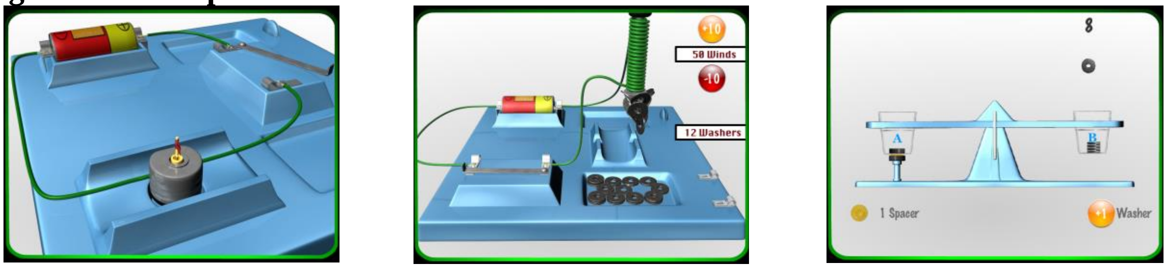
At any point that the student expresses a grasp of what a graph is, the tutor moves on to the next point.

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

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 …?"*

Figure 4 – Examples of Interactive Animations



a) Open and Close a Motor Circuit

b) Electromagnet simulation with changing variables

c) Breaking the Force Simulation

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 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; Mostoww & 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.

Figure 5 - Wizard screen

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:

- 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

Table 1 - Results for Speech Recognition

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 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 Track 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 format 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.

Figure 6 – Residual Gains

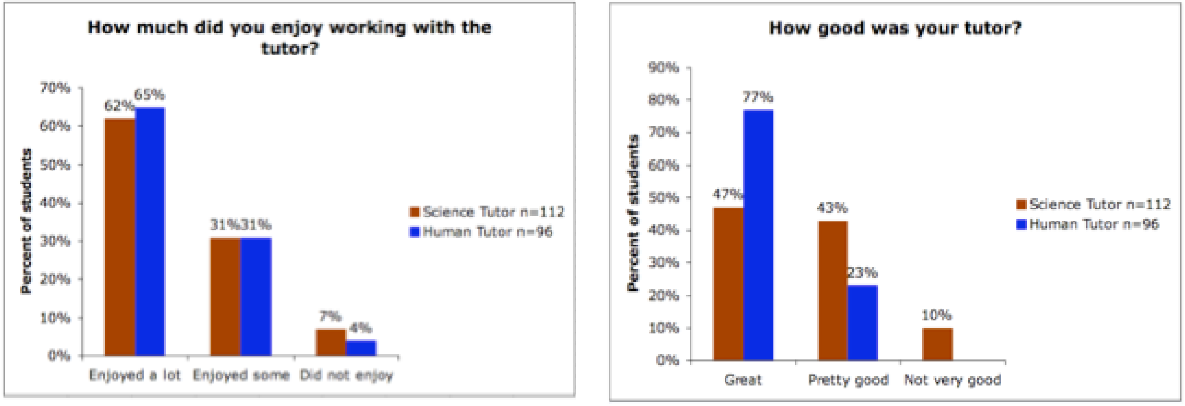
Research was conducted at schools with students from a large range of socioeconomic and ethnic backgrounds. Eighty-three (83) students received MyST tutoring, 69 were human tutored (both in 12 classrooms) and 1015 students in 50 classrooms in 20 schools received only classroom instruction and no tutoring. 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.

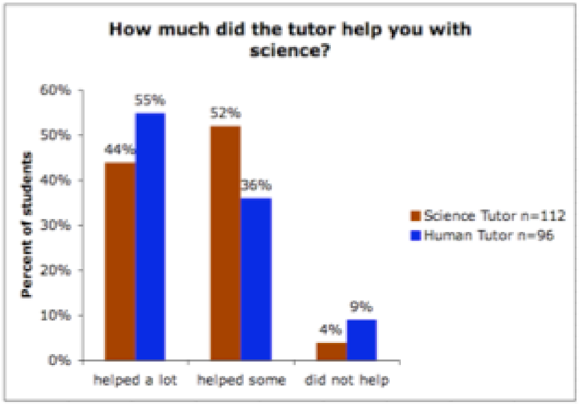
Direct comparisons of residual gain for the treat-ment groups (MyST and Human Tutored) showed no significant differences be-tween the two treatment groups with *t* = -1.14, d*f* = 150, *p* = 0.25. This supports Hypothesis 1, that learning grains from using MyST would be roughly similar to gains produced by human tutors. In the three-way comparison with the control group, MyST and human tutored groups had insig-nificantly 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, d*f* = (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* =.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, d*f* = 60, *p* = 0.014), as did the human-tutored group when compared with controls (*t* = 3, d*f* = 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.

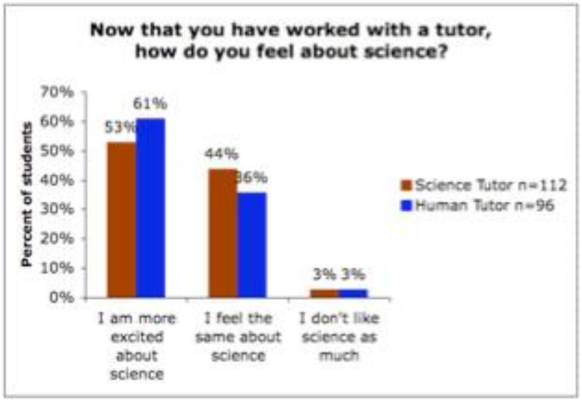
Is it Feasible to Integrate MyST-SLS into Elementary Science Curricula?

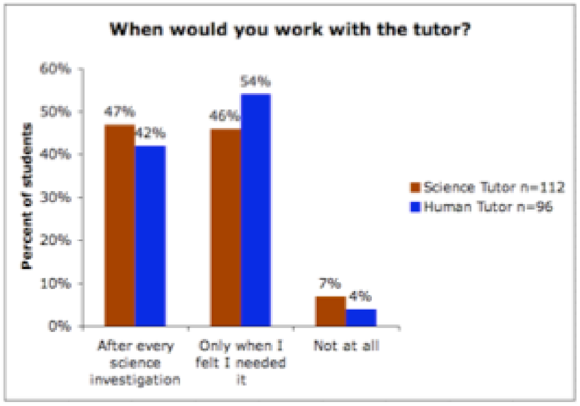
Figure 7 – Residual Gain by Pre-score

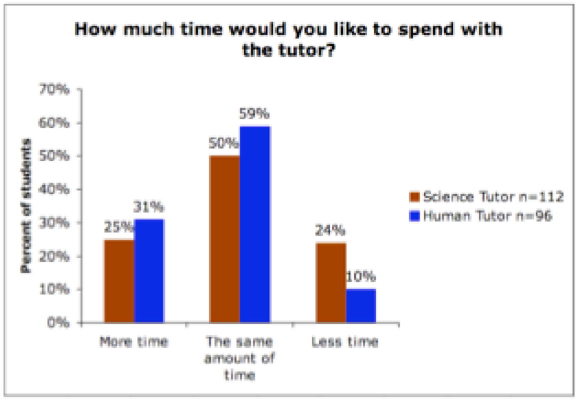
Figure 8 – Tutor vs Marni Survey Results











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 2010-2011assessment. 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

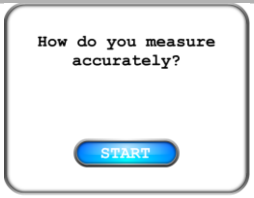
Title Screen: Each MyST-MP&D session began with a title screen that presented a deep reasoning question. In all cases, the printed question was read aloud by the virtual tutor. Examples included: What do magnets stick to? What is an electrical circuit? How can we measure length (volume, mass, temperature) 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



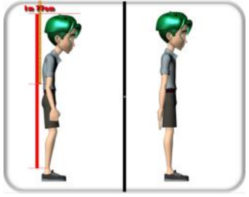
How do you measure accurately?

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

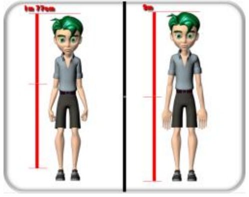
When ready they begin the CASUM tutorial by clicking on start.

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?"

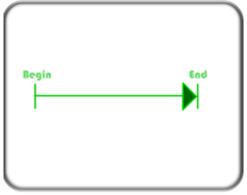
"How do you measure accurately?

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

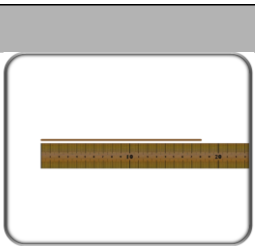
Suggested actions: Replay again, discuss what they notice going on about the quality of measurements that Jill is taking and why they are different. Then click on Continue. OR, just Continue to EXPLANATION.

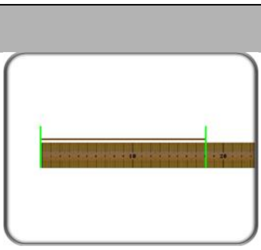
Figure 10

EXPLANATION



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

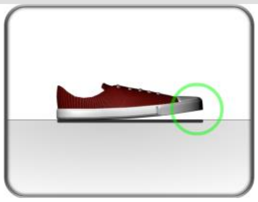




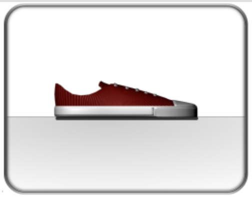
"It is also very important to make sure things are flat and lined up with your meter stick."

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?**

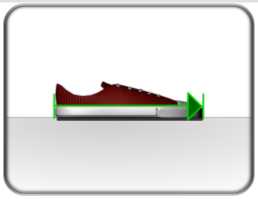
"Take this shoe for example."



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



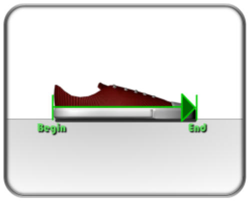
"In order to get a good measurement, we first have to make sure that the object



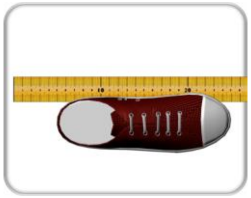
"Then we start our measurements at the back of the shoe

we are measuring is as flat as possible.”

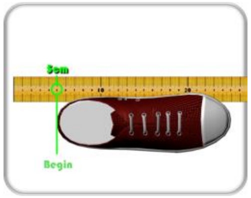
And measure all the way to the tip of the shoe.”



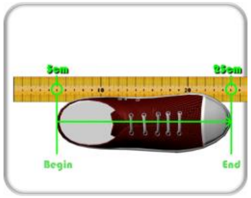
“These two spots are the beginning and end of our shoe”



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

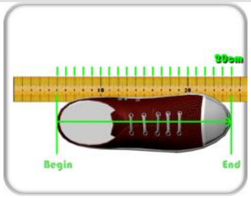


“When we look at the meter stick, we see that the back of our shoe is at the 5 cm mark…



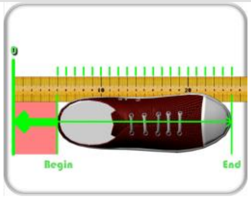
“…and the tip of our shoe is at the 25 centimeter mark. Does that mean our shoe is 25 cm long?”

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.

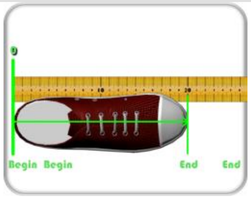


“ Well, when we actually count the centimeters starting at the back of the shoe and ending at the tip of our shoe….we count twenty centimeters, not twenty-five.”

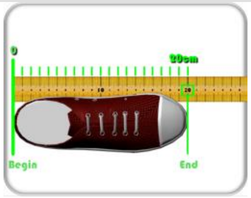
“Oh, instead of counting the units in-between, is there an easier way?”



“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.”



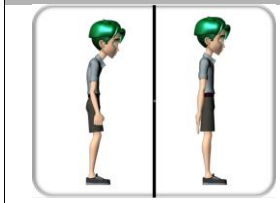
“And see what happens? Our shoe starts at the zero mark and ends at the twenty mark.”



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

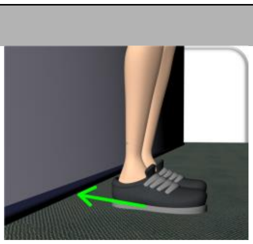
Figure 11

SOLUTION

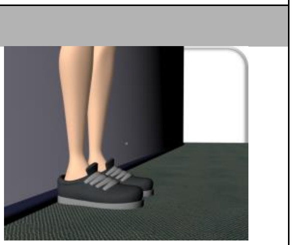


“Well Jill when you measured me you go two different measurements. Maybe how I was standing was part of the difference in those measurements.”

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

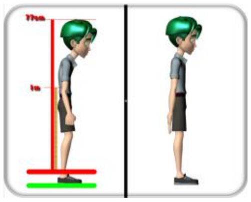


“And my first were not close to the wall, so I was not flat against the wall.”



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

“…so I was nice and straight all along the wall.”



“And remember the first time the meter stick was not down on floor by your feet. It was up

“So this time I'll measure the first meter…”

“…and mark it right here.”

	by your calf above your ankle. See, I started at the wrong place.”		
   “Then I’ll move the meter stick up and line it up with the mark.”	 **And finally I can measure the last length, which is a nother full meter.”**	 **“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”**	 **“So you are two meters tall; that’s pretty tall Jack.”**
	“Well, now that we figured it out…that was pretty easy.”		

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 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.

Table 3

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

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 a = .66 to a = .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.

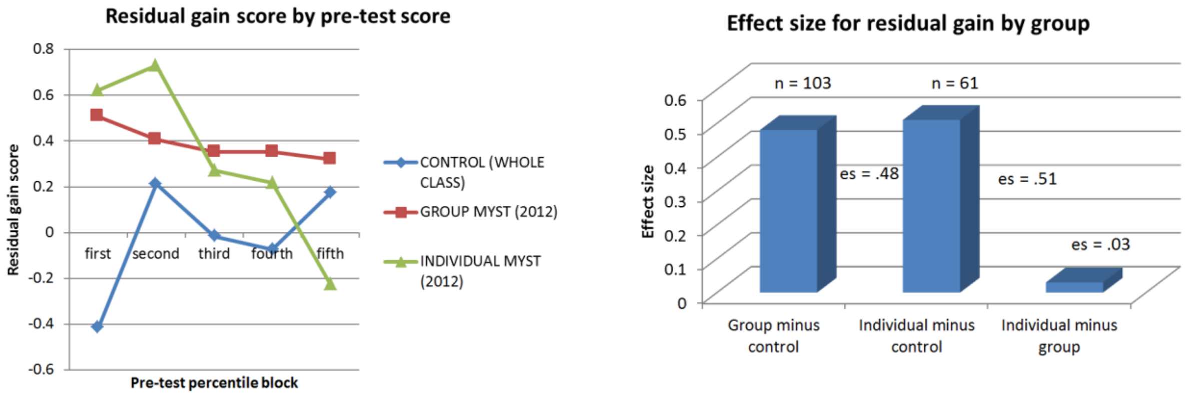
Results

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.
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.

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

(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

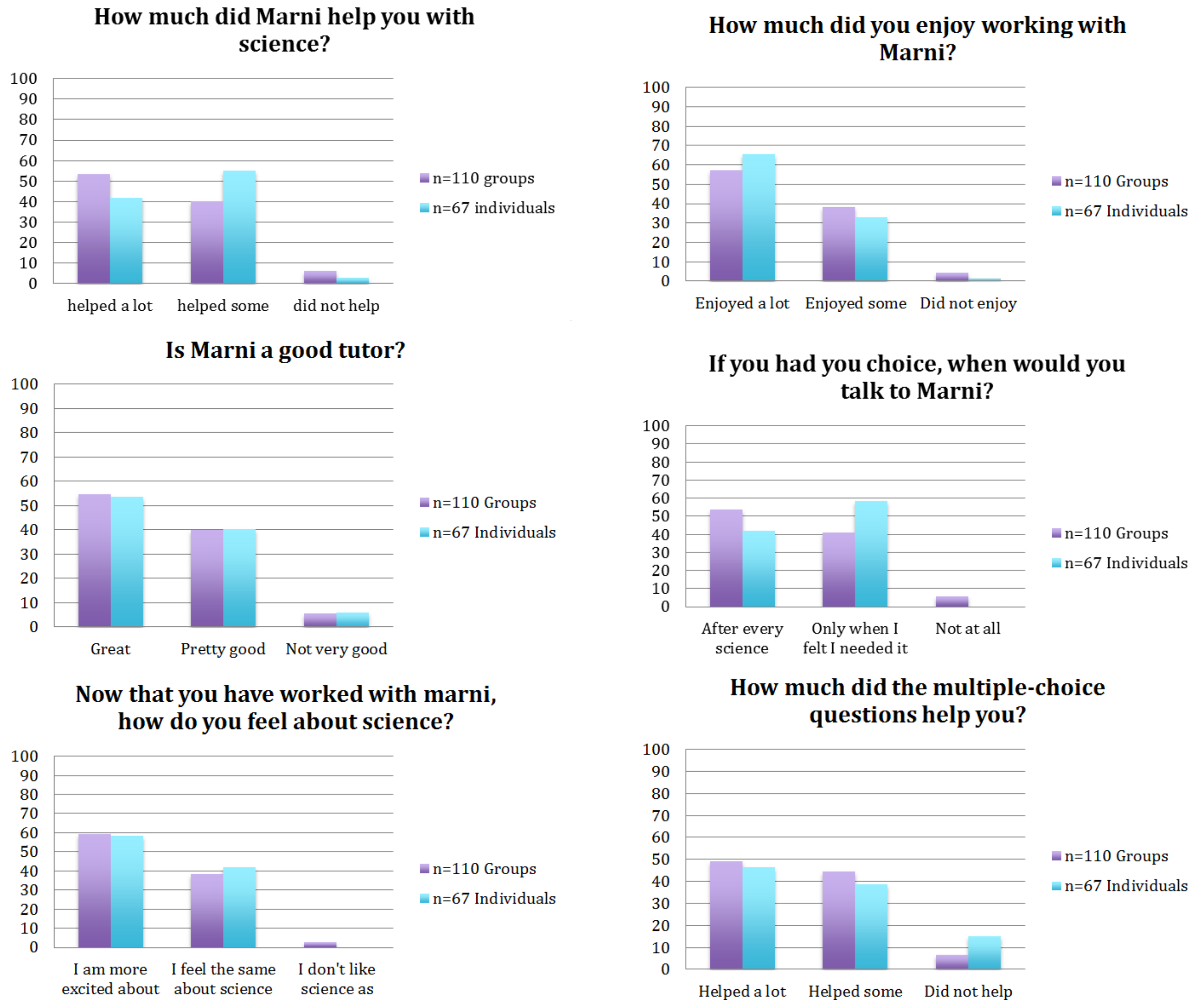
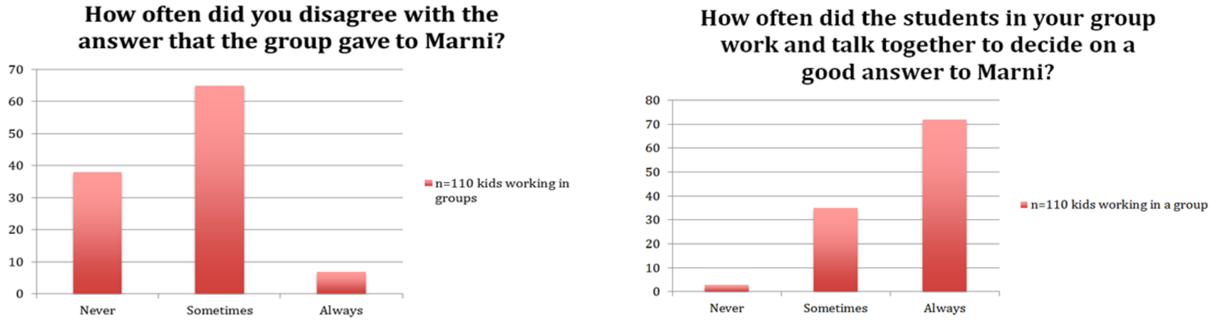
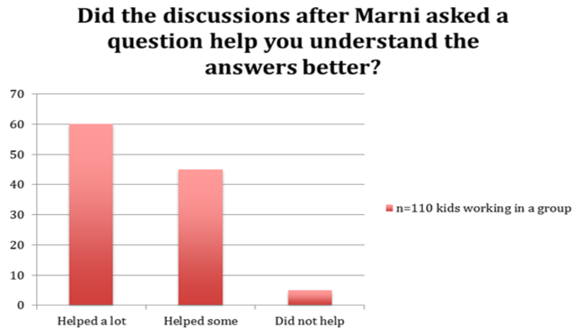


Figure 14 – Students in Small Groups Survey Results





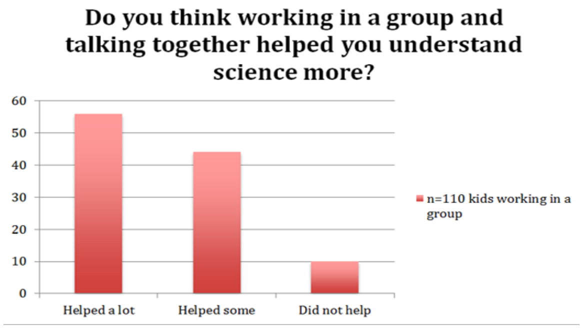
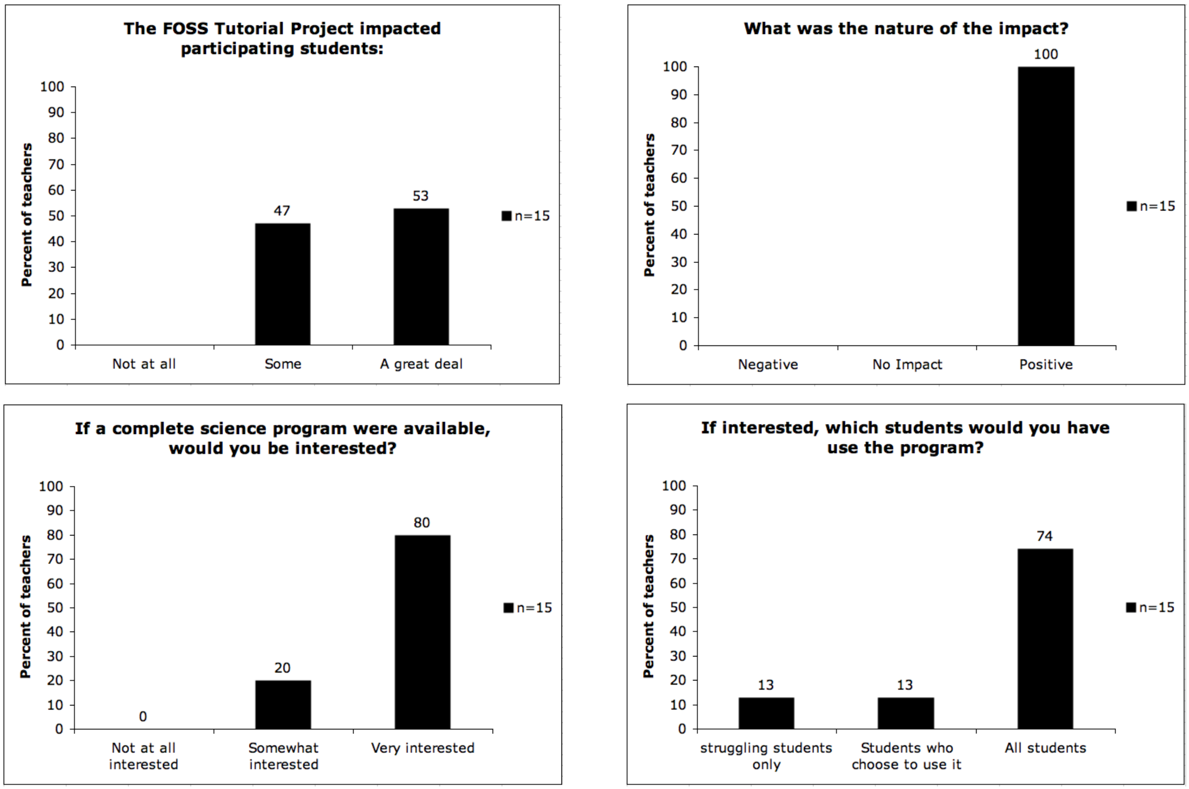
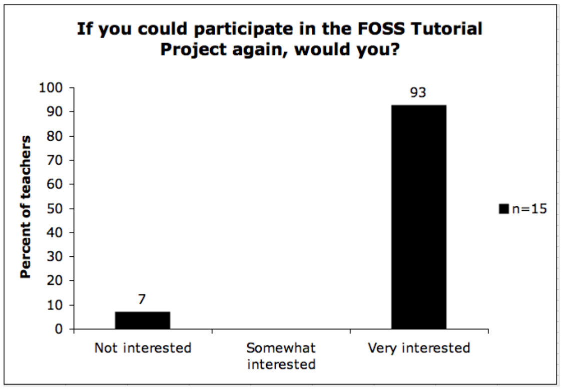
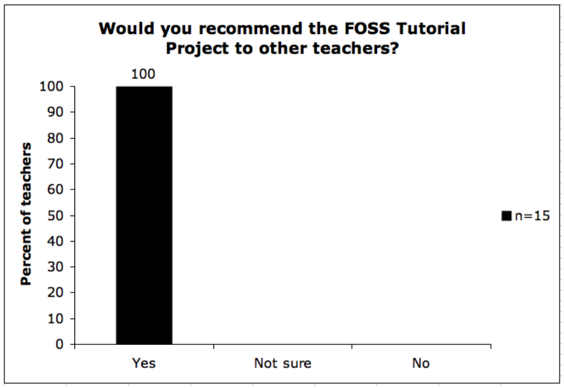


Figure 15: Combined Teacher Survey Results; MyST-SLS and MyST-MP&D




If you could participate in the FOSS Tutorial Project again, would you?
Percent of teachers
100
90
80
70
60
50
40
30
20
10
0
7
93
Not interested
Somewhat interested
Very interested
■ n=15


Would you recommend the FOSS Tutorial Project to other teachers?
Percent of teachers
100
90
80
70
60
50
40
30
20
10
0
100
Yes
Not sure
No
■ n=15

3. CASUM: Conversations About Science Using Media

CASUM is an intervention we designed to help teachers manage classroom conversations about science in which students build on each other's' ideas to construct explanations of science presented in teacher-controlled Flash animations. CASUM dialogs were inspired by two independent project components: a) the Flash animations that were developed for MyST-MP&D, and b) Questioning the Author, the approach developed by Isabel Beck and Margaret (Moddy) McKeown for teachers to manage conversations in which children talk about what the author is trying to communicate after reading students a passage of text.

We wondered why QtA, and classroom conversations in general, are not pervasive in US classrooms. Two large scale studies of classroom discourse, conducted by Nystrand and Gamoran (1991) and Nystrand, Gamoran, Kachur, and Prendergast (1997) indicated that teachers rarely ask students authentic questions that can lead to classroom discussions, and that extended conversations are practically nonexistent in elementary and middle school classrooms in the US.

These finding are puzzling, given that a) influential theorists and many educational researchers have both advocated and investigated benefits of collaborative discourse in classrooms, and b) meta-analyses of programs in which teachers learn how to facilitate classroom discourse provide strong evidence that well-managed classroom conversations improve learning and comprehension of texts (Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009; Soter et al., 2008).

During our interviews with teachers who used CASUM, several teachers independently shared that they did not feel comfortable classroom conversations in which students do most of the talking. This was especially true for the Magnetism and Electricity module, where conversations may lead to questions that teachers do not have the depth of knowledge to answer. This is not surprising, given that most teachers receive much less professional development related to science teaching, relative to reading and math. Teachers are comfortable teaching, they are less comfortable, and most have received little or no training, on how to facilitate classroom conversations.

After developing the sequence of activities in each tutorial dialog in MyST-MP&D, we realized that the Flash applications could be controlled by teachers, who could then ask students questions about the science exactly as Marni does in MyST dialogs. We therefore decided, with the enthusiastic support of Dr. Samantha Messier, the science curriculum director at BVSD, to investigate the feasibility of having teachers control the MyST-MP&D Flash applications while asking students questions designed to help them make connections with and build on each other's ideas to explain the science presented in the Flash applications. Two researchers associated with the MyST project, Cindy Buchenroth-Martin, who developed MyST dialogs and worked as a project tutor, and Liam Devine, who developed the Flash applications, worked with Ron Cole to implement the CASUM program. Cindy a) developed a Teacher Guide for each Flash application so teachers could review the Flash application, learn when to stop the animations and what questions to ask to initiate discussions, b) provided teachers' with professional development, and c) modeled one or two sessions for teachers before they conducted their own QtA dialogs with the Flash applications. Cindy also provided feedback to teachers following their first "solo" CASUM dialog.

During a CASUM dialog, the teacher initiates the discussion by displaying an illustration or animation. After a short interval during which students study the visual, she asks an open-ended question such as "What's going on here?" As students present their ideas, the teacher facilitates the conversation using a small number of effective “discussion moves” that help students clarify or expand their own or others’ explanations, make connections between ideas presented by different students, and challenge them to apply their explanations to new situations. At appropriate points in the discussion, the teacher refers to the illustration or animation while posing an authentic question that challenges students to apply their knowledge to a new situation. For example, if the animation shows electricity moving through a circuit, the teacher may ask her students: “What would happen if we flipped over the D-cell (battery) in the circuit?” After facilitating discussion on this topic, which may lead to alternative explanations or consensus, she says “let’s see what happens” and uses the mouse to click on the battery, which causes it to flip over and reconnect to the insulated wires on either side. Students are able to observe that the flow of electricity is reversed, and that a flag on top of motor changes direction. The teacher can then ask, “What just happened?” This experience with scientific conversation blends several dynamics widely recognized as cultivating scientific sophistication: it elicits the student’s thinking and draws the student into externalizing that thinking; it nurtures appreciation for alternative explanations and testable hypotheses; it emphasizes the role of observation and connecting observation to hypothesis; and it emphasizes the role of reflection and analysis after observation. The CASUM project relies on and researches this blend of dynamics as they are implemented through media.

Since the CASUM dialogs were based on the Flash applications in MyST-MP&D, each dialog was aligned to science concepts taught in classrooms using the FOSS program. We tested CASUM with both Measurement and Magnetism and Electricity and Measurement modules.

Sequence of Activities in Each CASUM Dialog

Title Screen: Each CASUM dialog begins with an important title screen (Appendix B, Figure 1a). The title of the CASUM session is presented as a question: Examples of CASUM titles are: What do magnets stick to? What is an electrical circuit? What is an electromagnet? The CASUM dialog is introduced with an authentic question because research has shown that presenting a student with an authentic or deep reasoning question before engaging in instruction or solving a problem improves learning (Driscoll et al., 2003; Gholson et al., 2009; Sullins et al., 2010).

Engaging Real-life Scenario: The first component of a CASUM dialog is a narrated animation that presents a real life scenario. For example, the scenario may show two characters (children, of different ethnicity and genders across different scenarios) talking about how their portable devices (e.g., cell phone, laptop, flashlight) can be turned on and off by pushing a button. How does this work? The scenario also introduces a problem that provides the basis for discussion. For example, Jack may remove a flashlight’s batteries, screw the cover back on, push the button, and wonder why it no longer works.

The Scenario introduces the science. It associates the science with materials and situations likely to be familiar to most or all of the students, and provides the basis for a discussion in which students can make connections between their own experiences and background knowledge about the science being learned. The Scenario discussion provides an opportunity for students to introduce vocabulary that the teacher can use to focus discussion (using a variety principled and substantive dialog moves discussed below): “So Jeannine just mentioned a D cell and electricity, what’s that all about? The teacher could also say: “I heard Jeannine talk about a D cell and electricity. Let’s look at this part of the movie. We see Jack pushing a button and the flashlight turns on. What does that have to do with a D cell or electricity?”

In sum, the Scenario provides the teacher with an opportunity 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 reflect on the Scenario's problem and its relationship to the deep reasoning question that introduced the CASUM session (Figure 9).

Explaining the Science: After students discuss the science problem, the teacher presents the students with a narrated multimedia explanation of the science. The design of these narrated animations is based on theory and research in multimedia learning, reviewed above. Each narrated animation is segmented into intervals, separated by pauses. Each interval presents a specific concept. The teacher is encouraged to stop the animation at these points and ask questions that lead to discussion of the concept, The sequence of concepts across these intervals are designed to build on each other, so the teacher can facilitate discussions that lead to complete and accurate explanations. 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. By designing narrated animations that present these component concepts sequentially, the teacher can pause the presentation after each segment to facilitate conversations that help students focus on the concept that was just seen and described, and then work with the teacher to construct an accurate explanation of each component concept, and then work together to integrate the concepts into a complete explanation (Figure 14). When students have demonstrated their understanding of these concepts, the Flash animation enables the teacher to test the students' ability to transfer their knowledge to a new problem. For example, the teacher can present a simulation of electricity flowing through a circuit, and ask: *What will happen if we flip the battery?*

Formative Assessment (MC Question): The CASUM dialog concludes with presentation of an authentic question that can be answered if students have achieved a deep understanding of the science. The question may be the same as the deep reasoning question that introduced the CASUM session or the question may be presented with a picture and require an answer that demonstrates application of the science knowledge to the situation shown in the picture. Following presentation of the question each student is asked to write an answer in their science notebook. When they have completed this task, students are presented with the question a second time, along with four response alternatives. The teacher then leads a discussion in which students discuss and defend the different response choices. After students have discussed their answers, each incorrect answer choice that students defended is reviewed and the correct choice is revealed and discussed. At the conclusion of the discussion, students may revise the answer they wrote in their science notebook.

CASUM Pilot Studies

Mandy' fourth grade classes

CASUM was first piloted in the spring of 2011 at a rural elementary school in Colorado. The teacher, Mandy, had three years' experience teaching science using FOSS. She taught two science classes with an average of fifteen children in a rural Colorado school. Her professional development consisted of an initial one hour meeting with Cindy Martin and Jeannine Moineau (also a dialog developer and project tutor). The group reviewed a set of Flash applications for Magnetism and Electricity (M&E), which Mandy would soon teach, and discussed how QtA could be used with the animations to engage students in discussions. Mandy was provided with Beck and McKeown's book on QtA *Improving Comprehension with Questioning the Author: A Fresh* *and Expanded View of a Powerful Approach* (Beck & McKeown, 2006). Prior to conducting her first CASUM dialog, Mandy watched Cindy and Jeannine conduct 3 CASUM conversations with her students. Mandy then took over, and was provided feedback on her first CASUM dialog. Mandy then carried out 11 CASUM dialogs with each of her 2 classes over 5 weeks.

Mandy reported that students were fully engaged in the conversations and in listening to and selecting response alternatives to the MC question. She greatly appreciated the ability to pace the conversations, and to play, pause and repeat the videos. She reported that the CASUM experience was a positive one; her students related well to the multimedia explanations, which strongly reinforced the scientific concepts they encountered in the classroom investigations. She noted that several students in her class, including a special education student with limited vocabulary, had great success acquiring and using authentic scientific vocabulary from the media and support from their classmates. Also, Mandy used CASUM in innovative ways: After watching and listening to the Explanations a few times, her students were asked to watch the animation without sound and explain what they were seeing. She asked students to go to the smart board to trace and explain the path of electricity in circuits. On two occasions Mandy had students use clickers to select response choices to the MC question; she reported that students really enjoyed using the clickers and seeing the responses made by all of the students. Mandy offered clear suggestions for improving CASUM which included a) improving the user interface to the Flash applications to navigate within each animation rather than only be able to repeat each segment, and b) shorten the animations; some animations were too long and presented too many concepts, which made the conversation more difficult.

Two CASUM pilot studies with English learners and students with special needs

We conducted two pilot studies of CASUM, in June 2011 and June 2012 with third graders who attended a Science Literacy Camp (SLC) organized by the Boulder Valley School District for ESL and Special Needs students. *All English learners who attended the SLC had low English language proficiency, based on test of English language proficiency administered by the Boulder Valley School District.* One week before the start of the SLC our project team held a 2 hour PD workshop with the 11 third grade science teachers. We discussed the goals of the pilot study, the scientific basis for CASUM dialogs, and conducted a CASUM dialog with teachers for "What is a Meter?" (The focus of the SLC was measurement; the students planned and planted a garden and conducted science investigations in the FOSS Measurement science module.) The June 2011 study involved 11 third grade classrooms in four schools, with approximately 15 students in each classroom. The June 2012 study involved 7 third grade classrooms in three schools. The students who were invited to attend the Science Literacy Camp consisted of English learners with low English language proficiency based on a standardized test administered by the BVSD school district, and special needs students. The 18 classrooms in the two studies taught the same science content, the same professional development for teachers, the identical Flash animations, and teachers were administered the same surveys at the end of the summer school. Based on these similarities, we combined the results of the two studies.

The day before the SLC began we visited each of the 18 participating classrooms to show teachers how to navigate within the Flash applications to present the media and pace discussions. Each teacher was given a teacher guide that summarized the Flash application, suggested logical stopping points, and provided examples of questions that could be asked to initiate discussions.

During the second day of classes in the SLC, a BLT project tutor skilled in QtA dialogs (through experience in the My Science Tutor project) conducted the first CASUM dialog in each of the 18classrooms. Each of the 11 teachers conducted 3 to 5 CASUM dialogs. Two or three days later, after students had completed a science investigation on measurement, teachers conducted their first QtA dialogs with students using the Flash application aligned to the science investigation and classroom instruction. A BLT project tutor attended each of these initial CASUM sessions, and provided feedback to the teacher. Teachers then conducted the remaining CASUM dialogs independently.

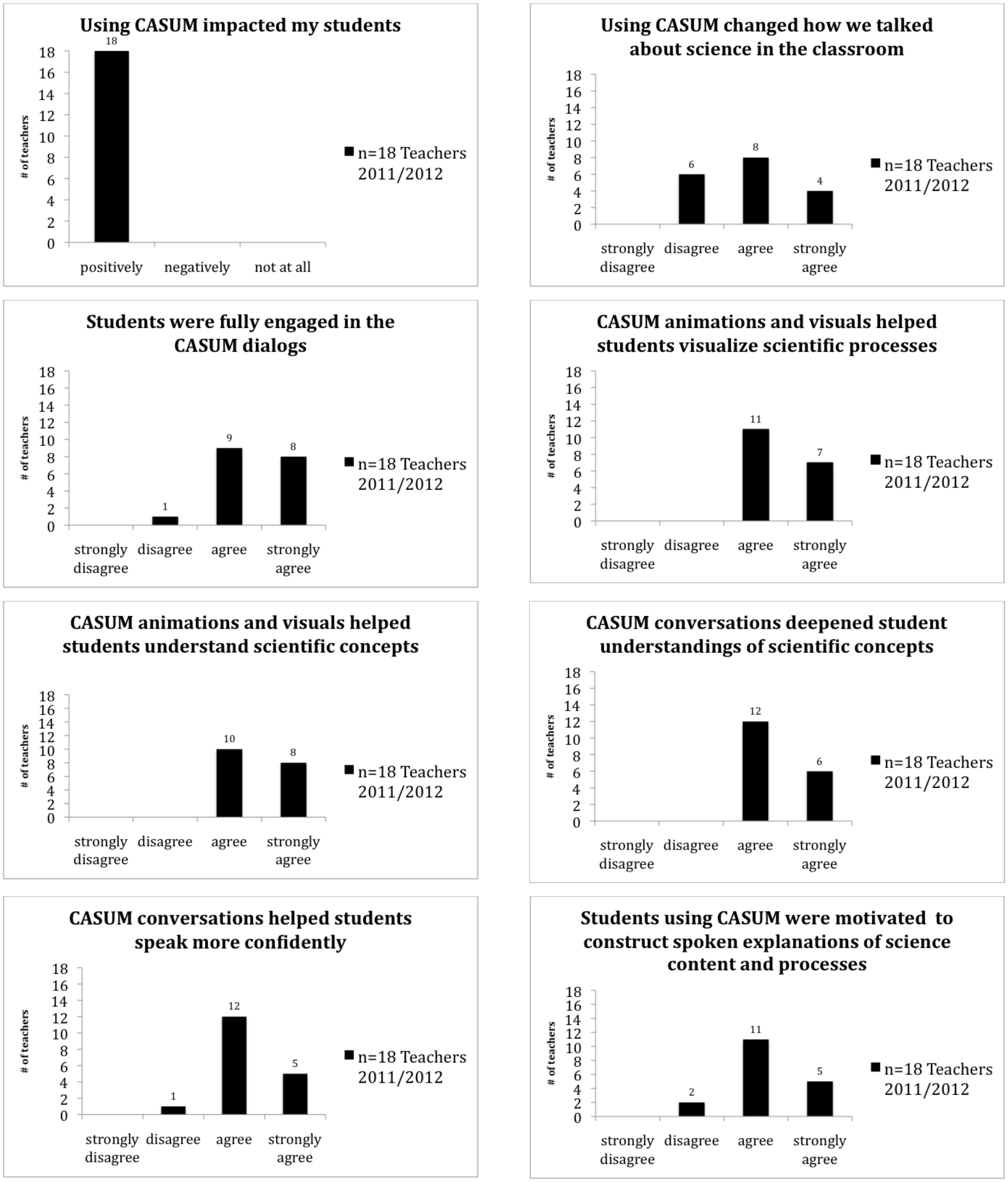
The CASUM dialogs for measurement featured two students, Jack and Jill. In the initial scenario, Jack and Jill set up the problem. For example, Jill offered Jack half of her milkshake. After pouring about half into Jack's glass (which was a different size than Jill's), Jack asks Jill how they could figure out how to share a drink and be sure each person had exactly the same amount. In the problem solving scenario, Jack and Jill work through the problem. Each of the 12 CASUM dialogs developed for Measurement featured Jack and Jill first introducing, and then working through a problem. The third grade students in the 18 classes became big fans of these characters.

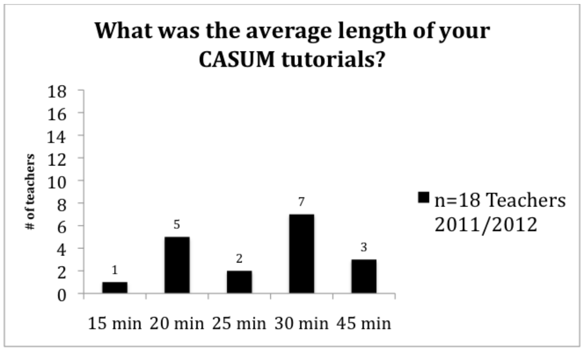
Results of the 2011 and 2012 CASUM SLC pilot studies. All 18 teachers responded to a 20 item survey and provided optional written comments following each question. A 4-point Likert scale was used to assess teachers' impressions of CASUM. The questionnaire included a set of positive statements, such as "CASUM conversations helped students speak more confidently" and teachers responded to the statement by selecting one of the 4 response categories: Strongly Agree, Agree, Disagree, and Strongly Disagree.

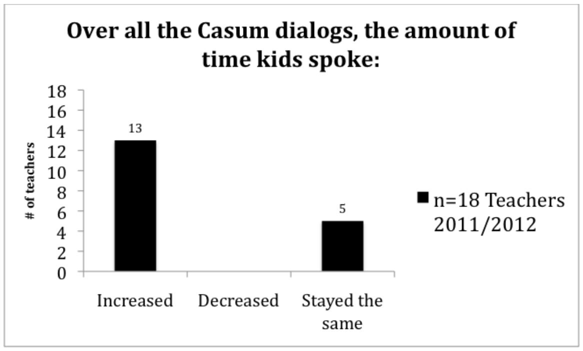
Teachers' responses to the survey questions are presented below as histograms that show the distribution of responses of the 18 teachers to each statement. It can be see that over 90% of all responses to questions about the perceived value and benefits of the program fell into the Strongly Agree and Agree categories. Teachers' written comments indicated that most were extremely enthusiastic about the potential of using CASUM dialogs, and would like to implement the treatment in their classrooms during the regular school year.

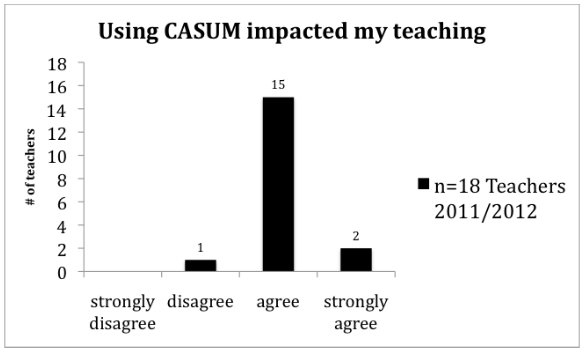
We have submitted two proposals to the IES and one to the NSF to develop and evaluate CASUM dialogs. These proposals were declined.

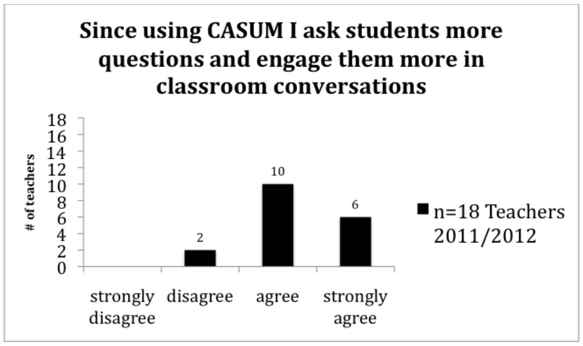
Figure 16 – Teachers Impressions of CASUM

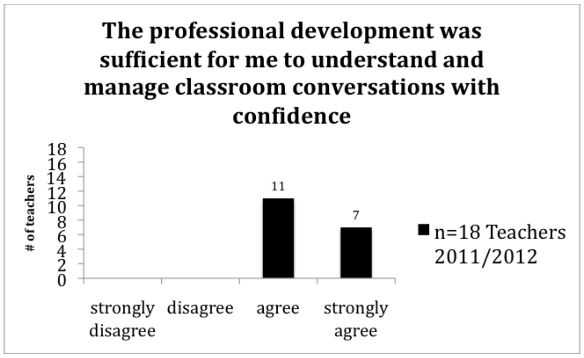


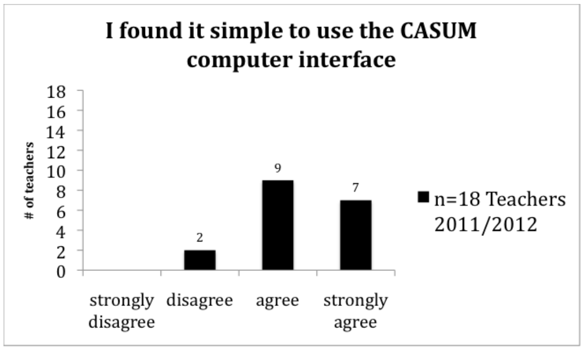

What was the average length of your CASUM tutorials?
# of teachers
18
16
14
12
10
8
6
4
2
0
1
5
2
7
3
15 min
20 min
25 min
30 min
45 min
n=18 Teachers 2011/2012

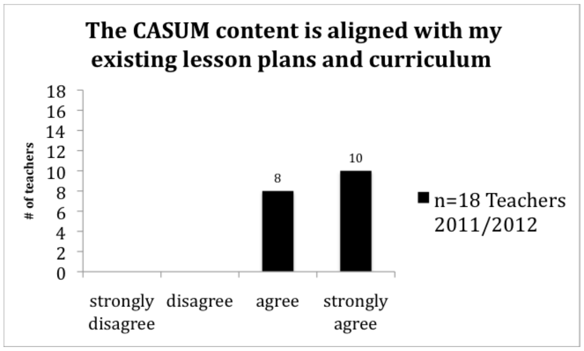

Over all the Casum dialogs, the amount of time kids spoke:
# of teachers
18
16
14
12
10
8
6
4
2
0
13
5
Increased
Decreased
Stayed the same
n=18 Teachers 2011/2012

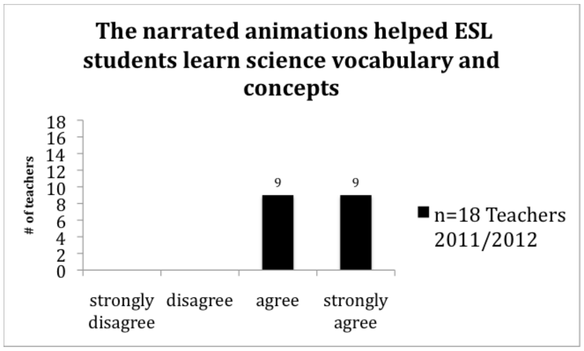

Using CASUM impacted my teaching
# of teachers
18
16
14
12
10
8
6
4
2
0
1
15
2
strongly disagree
disagree
agree
strongly agree
n=18 Teachers 2011/2012

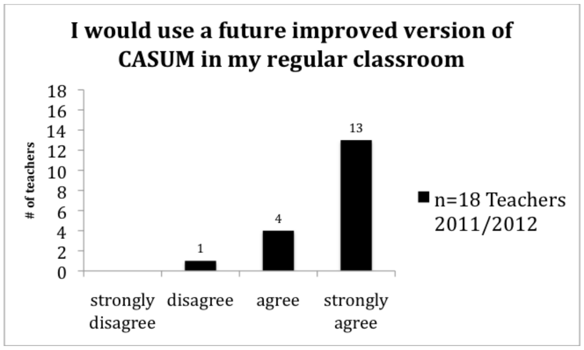

Since using CASUM I ask students more questions and engage them more in classroom conversations
# of teachers
18
16
14
12
10
8
6
4
2
0
2
10
6
strongly disagree
disagree
agree
strongly agree
n=18 Teachers 2011/2012

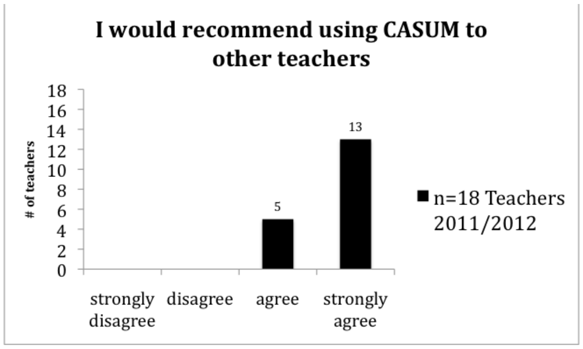

The professional development was sufficient for me to understand and manage classroom conversations with confidence
# of teachers
18
16
14
12
10
8
6
4
2
0
11
7
strongly disagree
disagree
agree
strongly agree
n=18 Teachers 2011/2012


I found it simple to use the CASUM computer interface
# of teachers
18
16
14
12
10
8
6
4
2
0
2
9
7
strongly disagree
disagree
agree
strongly agree
n=18 Teachers 2011/2012


The CASUM content is aligned with my existing lesson plans and curriculum
# of teachers
18
16
14
12
10
8
6
4
2
0
8
10
strongly disagree
disagree
agree
strongly agree
n=18 Teachers 2011/2012


The narrated animations helped ESL students learn science vocabulary and concepts
# of teachers
18
16
14
12
10
8
6
4
2
0
9
9
strongly disagree
disagree
agree
strongly agree
n=18 Teachers 2011/2012


I would use a future improved version of CASUM in my regular classroom
# of teachers
18
16
14
12
10
8
6
4
2
0
1
4
13
strongly disagree
disagree
agree
strongly agree
■ n=18 Teachers 2011/2012


I would recommend using CASUM to other teachers
# of teachers
18
16
14
12
10
8
6
4
2
0
5
13
strongly disagree
disagree
agree
strongly agree
■ n=18 Teachers 2011/2012

4. GROMINDS: Improving Science Learning and Reading Proficiency

During the fifth year of the project the PI and Co-PI of the MyST project worked with Dr. Julio Lopez-Ferraro, the DRK-12 Program Director, to receive a supplement to the DRK-12 MyST grant to participate in the Science Across Virtual Institutions (SAVI) project entitled *Innovations in Education and Learning* (http://innovationsforlearning.net). The SAVI consisted of 8 teams in the U.S. and 8 teams in Finland who worked together on 8 different projects. The common theme across all projects is **engagement.** The different projects all seek to gain new knowledge about how student engagement in learning tasks can be measured and increased to improve learning. Our project, called GROMINDS, partners researchers_at Boulder Language Technologies (Ron Cole, Eric Borts), Southern Methodist University (SMU; Doris Baker) and Pepperdine University (Eric Hamilton) with a team of researchers at AGORA center at the University of Jyväskylä in Jyväskylä Finland (Heikki Lyytinen, Ulla Richardson, Jarkko Hautala, ***and Aleksi*** Keurulainen). A major focus of this collaboration was development of **MindStars Books**, developed at BLT, and **Graphogame**, developed at the AGORA center, for use in both U.S. and Finnish elementary schools, . Participation by the U.S. team was supported by a one year supplement (September 2012-August 2013) to NSF grant 0733323: *Collaborative Research: Improving Science Learning in Inquiry-based Systems.*

Specific Aims of the GROMINDS Project

GROMINDS was planned as a two year SAVI project with the following objectives:

1. Design and test an initial prototype of the **MindStars System Architecture** (MSA) to provide 24/7 access to the resulting MindStars Science Books. The MSA will be configured as a set of Web services that support both Graphogame services (via servers in Finland) and MindStars Science Books. The goal is to provide web-based learning tools that are platform-independent, so they can be used on desktops, laptops, notebooks and mobile devices. In addition, we envision MSA as a free set of tools and technologies that researchers worldwide can use to replicate and extend our intelligent tutoring systems. We will provide the systems, documentation and tutorials so that others can develop and integrate BLT's speech recognition, natural language processing, dialog modeling and character animation technologies into tutorial dialogs in MindStars Books. These tools and technologies can also be used to develop speech corpora for training speech recognizers in different languages, so that Marni can interact with children in any language for which a recognizer has been developed.

2. Develop and investigate how English and Spanish versions of Graphogame (https://Graphogame.com/) can be used to help English learners and other struggling readers learn to read words in Spanish and English texts accurately and effortlessly. Graphogame has been shown to provide a powerful and flexible tool for helping students acquire *word reading automaticity,* a foundational reading skill that is necessary for reading texts fluently and with good comprehension. When students achieve word reading automaticity, they are able to devote their cognitive resources to making sense of the text; they change from students who are learning to read to students who are reading to learn (LaBerge & Samuels, 1974; Perfetti, 1985) Graphogame has been demonstrated to be a highly effective tool that can be integrated into classroom instruction to significantly improve children's ability to acquire the sound-letter correspondences needed to recognize words accurately and automatically, with long term benefits demonstrated for students in different countries and languages. Our research will be the first to extend Graphogame to U.S.

students, and investigate hypothesized benefits of combining Spanish and English versions of Graphogame to optimize acquisition of *word reading automaticity*, which predicts children's future reading fluency and comprehension.

MindStars Books

The Vision: MindStar Books represents an imaginative new generation of intelligent tutoring systems in science and in reading. Our vision builds on prior generations of intelligent tutoring systems, including significant foundational work carried out by the project team under NSF and IES support. We seek great strides in the quest to immerse students more effectively in multimedia learning activities in which they are challenged, motivated and empowered to acquire the knowledge and skills to learn science.

MindStar Books are thus designed to scaffold effective science learning with the following four aims: 1) They will enable students, especially including English language learners, to acquire the prerequisite vocabulary and concepts to listen to and understand science texts that are read aloud to them by a virtual tutor while they view illustrations that help them visualize the science being explained. 2) They will assess students' understanding of the science through spoken presentation of deep reasoning questions, challenging answer choices representing common misconceptions, and immediate formative feedback on their answer choices. and 4) MindStar Books will engage students in activities that lead to accurate, fluent and expressive reading of grade-level texts; skills that correlate highly with reading comprehension and future reading success (Baker et al., 2008; Fuchs, Fuchs, Hosp, & Jenkins, 2001; LaBerge & Samuels, 1974; Perfetti, 1985; Reynolds, 2000; Samuels, 1997; Stanovich, 2000). These are important and exciting aims based on prior research and development, and they are within grasp.

Scientific Foundations: MindStars Books are based on theory and evidence indicating that a student's ability to read and understand a text– their reading comprehension ability– consists of two component skills: *listening comprehension* and *word reading automaticity.* Listening comprehension is an individual's ability to listen to a text and answer spoken questions about it. Reading fluency is the ability to recognize words accurately and effortlessly. Research shows that students' reading comprehension abilities can be accurately predicted by independent measures of their listening comprehension skills and their ability to recognize words accurately and rapidly(Gough, Hoover, & Patterson, 1996; Gough & Tunmer, 1986; Hoover & Gough, 1990). MindStar Books are designed to help students develop these two essential skills.

MindStars Books Development Efforts

MindStars Books Toolkit: Research conducted during the GROMINDS project resulted in development of the MindStar Books Toolkit, an authoring environment for building, testing and publishing the MSBs. To date, 8 complete books have been developed and tested in schools. An additional eight books are under development. These books are aligned to Next Generation Science Standards, (NGSS), Colorado standards, Texas standards, and FOSS learning objectives for elementary school life science. Our near-term goal is to develop a complete sequence of books using the MSB toolkit for life science that are leveled to grades K-5. All but the kindergarten books will incorporate oral reading fluency training following listening comprehension training.

Design and Organization of MindStars Books

The MindStars Books Toolkit: The MindStars Books Toolkit was developed to provide an easy to use authoring environment for developing the listening comprehension activities in MS Books, and publishing the book in a library. The tool enables an author to (a) type in each sentence Marni will say, (b) record the sentences in English and record the Spanish translation of each sentence, (c) select a picture that will be presented with each narrated sentence (portions of pictures are highlighted using Photoshop), (d) include optional sound files into the narration, (e) design one or more multiple choice questions, with optional illustrations, that are presented after the page has been narrated, and (f) record the questions and answer choices in both English and Spanish. Once the listening comprehension activities have been developed, the oral reading fluency training activities, which follow listening comprehension, are generated automatically, using the text that is narrated by Marni during listening comprehension training. In May 2013, BLT hosted a workshop in which 7 research staff from BLT and SMU and a primary school teacher learned to use the authoring tools to create new books.

Listening Comprehension: In MS Books, Marni narrates each page of a science text while the student views illustrations that help them visualize the science. The narration is self-paced in alignment with research that indicates that self-paced presentations improve learning (Barker, 2003; Cole, Halpern, et al., 2007; Cole et al., 2003) . Students can stop and resume the narration after Marni speaks each sentence, and have Marni repeat the sentence in English or say a Spanish translation of the sentence. After listening to one or more pages of text, Marni presents students with multiple choice questions (MCQs) to assess their understanding of the vocabulary and concepts. These are deep reasoning questions with challenging answer choices that represent common misconceptions. Students can listen to the question and answer choices either in English or in Spanish as often as they like. After selecting an answer, the student receives immediate feedback about the answer they selected. Marni provides positive feedback to a correct answer. If the student selects an incorrect answer choice, Marni scaffolds learning by providing a hint; e.g., that spider has 8 legs, so it can’t be an insect. After two tries, the correct answer is presented to the student, along with an explanation as to why the answer is correct. We note that during listening comprehension activities, words are not presented on the page, as the goal is to have students listen carefully while viewing illustrations; research indicates that printed words can distract the student’s attention from the illustrations and reduce learning (Cole, Wise, & Vuuren, 2007) .

Oral Reading Fluency (ORF): ORF practice and training occurs immediately after the listening comprehension activities are completed; that is, after all pages of the science text have been narrated to the student and MC questions have completed. The goal of the ORF training is to help students learn to read grade level science texts accurately and fluently; oral reading fluency has been demonstrated to be a strong predictor of reading comprehension and later reading proficiency (LaBerge & Samuels, 1974; Perfetti, 1985; Ward et al., 2011; Ward et al., 2013). Fluency training occurs through repeated reading of each page of the science text. The student is presented with the first page of the text, with each sentence displayed on the page. The student can choose to practice reading the text, with support from Marni, before reading it independently. During practice, the student can listen to Marni read an entire sentence, or pronounce individual words in a sentence. The student can record themselves reading these sentences or words and play back their recordings to compare their reading with Marni’s. During playback of their recordings, each word is highlighted on the page as it spoken by the student. English learners can listen to Marni read a translation of the sentence in Spanish. When the student has finished practicing, they click an icon to read the page independently. Immediately after reading the page, the student receives feedback on the number of words they read correctly (out of the total number of words on the page), and their reading rate (relative to Marni's natural reading rate). The MindStar book highlights words that the speech recognizer scored as misread or skipped, so the student can practice reading these words and sentences. Repeating readings of the page, with practice before each reading and feedback on the student's reading performance immediately after independent reading, continues until the student achieves a criterion level of oral reading performance (90% word reading accuracy, reading speed within 10% of Marni's) or after three independent readings. Repeated reading of texts with feedback and practice following each reading has been shown to be a powerful tool to improve reading fluency, which correlates highly with reading comprehension (Baker et al., 2008; Fuchs et al., 2001; LaBerge & Samuels, 1974; Perfetti, 1985; Reynolds, 2000; Samuels, 1997; Stanovich, 2000).

MindStars Books Pilot Study

An initial set of eight books that supported listening comprehension activities in English, with Spanish translations of Marni's speech, which could be invoked by users, was tested in Kindergarten and first grade classrooms during May, 2013. The books were based on life science themes (e.g., Insects, Life Cycle of the Monarch Butterfly, What do Animals Need to Live?). Students were able to start and stop the (self-paced) multimedia presentation at any point, and repeat the entire presentation if they desired. They could also have Marni repeat questions and answer choices as often as they liked before choosing an answer. When a correct answer was chosen, Marni provided positive feedback to the student, and often expanded on the correct answer. When students chose an incorrect answer, Marni provided a hint (e.g., that spider has 8 legs, so it isn't an insect), and asked the student to choose again. If the second choice was incorrect, Marni explained why the choice was incorrect, and provided the student with the correct answer. English learners who spoke Spanish as their first language also listened to Marni in English during these activities, but had the option of clicking on an icon to hear Marni produce a Spanish translation.

The results of the pilot study provided initial evidence that children were highly engaged in using the books and that they were effective in helping students learn science vocabulary and concepts. Eight kindergarten and first grade students interacted with Marni in 6 to 8 different books. All together, the students were presented with 301 multiple choice questions. Across all students, 273 (90.7%) were answered correctly on their first choice. An additional 26 (8.3%) were answered correctly on students' second choice. While these results are encouraging, additional research is needed to determine if students will retain the knowledge acquired in the books, or will be able to transfer it to new contexts.

One English learner who used the books consistently invoked Spanish translations of prompts by Marni, and these occurred most often during their review of answer choices to multiple choice questions. The student who invoked Spanish translation most often—over 90% of the time during MCQs—was proficient in English. This student shared with us that he enjoyed listening to Marni explain science, ask questions and read answer choices in both English and Spanish because it helped him understand the science.

Figure 17
– MindStar Books UI

Research conducted during the SAVI project resulted in development of the MindStar Books Toolkit, an authoring environment for building, testing and publishing the MSBs. To data 8 complete books have been developed and published. An additional eight books are under development. These books are aligned to Next Generation Science Standards, (NGSS), Colorado standards, Texas standards, and FOSS learning objectives for elementary school life science. Our goal is to develop a sequence of books that are leveled to grades K-5. All but the kindergarten books will incorporate oral reading fluency training following listening comprehension training.

In addition, collaboration with researchers at the University of Jyvaskyla *resulted in Finnish versions of the eight MSBS developed thus far.* In these books, Marni speaks Finnish, and can produce English translations of Finnish prompts. The development of the Finnish speech recognition system, by the Finnish research team, was an important outcome of the collaboration. The Finnish research team is planning to evaluate the Finnish MSBs in third grade classrooms in Jyvaskyla schools in January, 2014. These books will include an initial version of the oral reading fluency component.

MindStars Books: Hopes for the Future

Development and evaluation of MindStars Books was planned as a two year project with our colleagues at the *University of Jyväskylä. In the context of the NSF SAVI project,* the *Finnish* Funding *Agency* for Technology and Innovation (TEKES), awarded two year research grants to each of the eight Finnish teams. All of the U.S. teams except BLT are funded for the second year of the SAVI project. BLT is actively exploring options for continuing this project. BLT submitted an ambitious proposal to the NSF Cyberlearning program to develop and assess MS Books. The proposal received strong reviews and was judged as competitive, but it was declined. In December 2013 BLT submitted a proposal to the NSF SBIR program to investigate the feasibility of commercializing the MS Books.

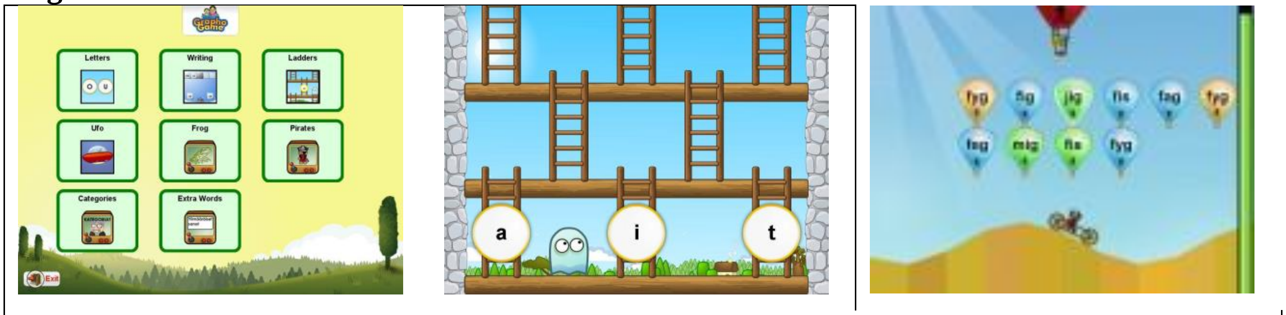
Graphogame

Graphogame was designed to help children acquire foundational reading skills that are taught in first, second and third grades. Many students in U.S. schools fail to acquire the foundational reading skills to read texts fluently and comprehend them. Students who do not learn to read words accurately and automatically have a difficult time reading with comprehension, and are cannot learn from texts as well as their more reading-proficient peers. One of the most fundamental skills in reading is the ability to read words accurately and effortlessly—i.e., acquiring word level automaticity. Graphogame (GG) was designed to enable children to learn sound-letter correspondences in different languages so they can learn to read words accurately and automatically

Graphogame is an online educational software for training and assessing students' reading skills. It was developed by the University of Jyväskylä (Finland) and the Niilo Mäki Foundation (Jyväskylä, Finland). Graphogame is currently available for research purposes outside of Finland. In Finland, the game is called Ekapeli, and it is delivered online free of charge to all learners (see http://info.Graphogame.com/).

GG consists of a number of different activities that build on each other to teach these skills. For example, in one exercise, GG presents young learners with two or more letters (or syllables or words) that fall from the top to the bottom of the screen (on a tablet computer in our research) while an auditory stimulus corresponding to the stimuli is presented to their ears. Their task is to choose (by pointing or shooting) the letter(s) corresponding to the spoken stimulus before the letters reach the bottom of the screen and disappear. Children’s responses are recorded and the system adapts to each individual student’s performance to optimize learning by adjusting the speed of the falling letters or the number of repetitions of the stimuli before introducing new choices.

Figure 18



GraphoGame Screenshots

Graphogame has demonstrated success in helping students achieve word reading automaticity, an essential component of reading with comprehension. The game is very intuitive and all children appear to grasp it immediately. GG has been tested successfully by more than 200,000 children in Finland, Zambia, Great Britain, and Chile (Brem et al., 2010; Kujala, Lovio, Halttunen, Lyytinen, & Näätänen, 2012; Kyle, Kujala, Richardson, Lyytinen, & Goswami, 2013; Ojanen, Kujala, Richardson, & Lyytinen, 2013).

The aim of our research using Graphogame-English and Graphogame-Spanish was to optimize learning of word-level decoding skills by English-only students and English learners, so they can recognize English words in texts accurately and automatically. As we have noted, developing word recognition automaticity is a critical skill that children must acquire to read texts fluently and comprehend them.

Research Questions

The aim of the Graphogame study is to provide initial evidence of the usability, feasibility and promise of Graphogame to increase the decoding skills of first grade students in English and in Spanish. Specifically, we aimed at answering the following research questions.

1. Did students who participated in the Graphogame-English condition perform better on letter sound, word reading, and oral reading fluency (ORF) at posttest in English compared to students who did not play Graphogame taking pretest scores into account?
2. Did students who participated in the Graphogame-Spanish condition perform better on letter sound, word reading, and ORF at posttest in Spanish and English compared to students who did not play Graphogame taking pretest scores into account?
3. Did students find Graphogame engaging?
4. Did teachers find Graphogame easy to incorporate into their regular classroom routines?

We also included secondary questions that would help us identify potential moderator and mediator variables that could have affected results such as (a) individual differences in decoding skills at pretest, (b) classroom random assignment, and (c) amount of time playing the game.

GG Summative Evaluation

Research Design

We randomly assigned classrooms within Dallas Texas schools to playing Graphogame 10 minutes/day for 16 weeks versus business as usual instruction. The amount of time per day was determined based on previous studies on the efficacy of Graphogame (Kyle et al. 2013), and on the amount of time it was feasible for teachers to use the game during the English Language Arts block.

Prior to the beginning of the intervention, we met with teachers to explain the project and obtain their consent to participate. A total of 14 first grade classrooms in four schools agreed to participate. We randomly assigned these classrooms within schools to either a treatment condition or a control condition. Given that in one school there was only one English-only classroom and one bilingual classroom, both classrooms were in the treatment group after the random assignment.

Students in the treatment classrooms received headphones and general instructions on how to log into the game. Teachers used a timer to control the amount of time students used the game and to provide all students an opportunity to play the game. Students in the control classroom did not receive any opportunity to play the game during the 16-weeks of the study. However, after that time, teachers in the control condition were also eligible to use the game in their classrooms, if they wanted to.

All students in each of the 14 classrooms were eligible to participate in the study. Five English-only classrooms and three Spanish-English classrooms were in the treatment group and four English-only classrooms and two Spanish-English classrooms were in the control group. Spanish-English classrooms provided reading instruction in Spanish and English every other day to students who spoke Spanish at home as determined by a parent survey.

Participants

Schools: Four schools in a large metropolitan area in Texas agreed to participate. Three of the four schools were located in a high poverty neighborhood. The number of students in each school ranged from 261 to 620. The fourth school served about 1,300 students in Grades K-12. Table 1 presents information about ethnicity, percentage of students on free and reduced lunch and percentage of students who received special education services at each school. Table 1 in Appendix A2 presents student demographics by school.

Students: 268 students participated in the Graphogame study. Fifty-four percent were male, 66% were economically disadvantaged, and 28% had limited English proficiency. Eighty-three students (31%) of the 268 were Spanish-native speakers and therefore were receiving Spanish and English reading instruction in a one-way dual language model. These students played Graphogame in Spanish.

Reading Instruction: In all schools students received approximately 60 minutes of daily reading instruction following the Journeys core reading program (Harcourt, 2010). The program has also a Spanish version, Senderos (HM, 2010), that was used in the Spanish language arts classroom. Teachers taught the reading programs following the specified scope and sequence, and the instructional guidelines suggested by the publisher. Students who were struggling with English also received supplemental instruction from their classroom teachers in an after-school care environment based on teacher recommendations. Supplemental instruction often included specific decoding and vocabulary lessons.

Professional Development: We trained teachers for half a day on the implementation of Graphogame. Teachers logged on to the game website, played the game, learned the different levels that students would progress through, and received detailed information about the best time to play the game. Teachers asked clarifying questions and received feedback and assistance from the research staff. Teachers and students also received technical support and Graphogame assistance by a research assistant in the beginning of the intervention. Research staff was also on call for questions from teachers during the 16-week study.

American English and American Spanish Graphogame Study Versions

Two different versions of the same Graphogame software platform were used in this study, the Graphogame US Rime English and the Graphogame US Spanish. In both versions the functionality and visuals of the central tasks are the same (i.e., the players are trained to connect spoken sounds to their corresponding written formats in either single letters or a longer sequence of letters). The task of the player is to first listen to a spoken sound, and then select the correct corresponding written target from several alternatives by using a computer mouse. After an incorrect selection, the game immediately shows the correct correspondence for the presented sound. The game never gives any negative feedback on the performance but instead provides the correct answer and/or provides positive feedback on the correct selections. Each game level is relatively short, lasting on average from one minute to three minutes.

Both game versions adapt to the performance of the player. In other words, the game constantly keeps logs and according to the performance in each particular trial, the program provides learning material in the following trials or levels that allow the player to achieve an average of 80% correct items. In addition, similar game levels are presented in several different types of graphic settings allowing students to practice the same task hundreds of times. The design of the game keeps students interested in playing the game and increases their engagement. Below we describe some of the differences between the English and the Spanish US version of Graphogame.

Graphogame English**:** The original version of Graphogame was developed with the support of British reading researchers (Kyle et al., 2013) and then modified by replacing the sounds of the letters in British English with the sounds of a general American accent recorded by a native American English speaker. The reward system in the Graphogame US English Rime game allows the player to select virtual stickers that can then be placed in a sticker book. The learning content in both of the game versions moves from small units to larger units. The learning content is organized into streams with several levels that explicitly instruct learners on orthographic rime units (Kyle et al., 2013). First, the game introduces the specific grapheme-phoneme connections (GPCs) that form specific rhyming word families. Second, players get to play with rime units thus, providing opportunities for players to reinforce their GPCs skills facilitating the recognition of psycholinguistically relevant reading and spelling units in English. The last step in a stream is to play with words that contain the rime units of the previous levels.). For example, in stream 1, a small set of seven single phonemes and graphemes is introduced (C, S, A, T, P, I, N). Students are then told to put these sounds together to make rime units. Once this is accomplished, students are told to put another sound in front of the rime units presented. The order in which the rime units are introduced is based on the phonological neighborhood density of the rime units according to a constructed database (deCara & Goswami, 2002)..

Graphogame Spanish: The original Spanish US game was developed for an intervention study in Chile. Sounds of letters that seemed unclear or confusing were rerecorded or deleted for the US Spanish version. Learning content in the US Spanish version is organized into dimensions that are listed by the order of difficulty. The game adaptation algorithm works in two phases (i.e., on the accuracy scores in the earlier play sessions, and in the last ten trials of the current run within each content type). The game levels are placed on a map where the player character moves. The performance in each level is rewarded with points that can be used for gaining more power, speed, style or protection for the player character. In addition, the player character gains game tokens that can be used to obtain various objects such as head wear, clothes, and toys.

Measures

Students receiving English instruction only were assessed with three measures. One measure was part of the GG game, the Graphogame US English assessment, and the other two were part of the Dynamic Indicators of Beginning Early Literacy Skills (DIBELS) (Good & Kaminski, 2002) system. Bilingual students receiving Spanish and English reading instruction were assessed in both, English and Spanish. In addition to the three English assessments, bilingual students were also assessed with one Graphogame Spanish assessment, and two measures of the Indicadores Dinámicos del Exito en la Lectura system (IDEL) (Baker, Good, Mross, et al., 2006). Appendix A provides detailed information about the English and Spanish measures.

Student Survey: The purpose of the student survey was to determine student self-reporting of their engagement and ease of use of the game. The student survey was adapted from a previous survey and consisted of 8 questions such as *How much did you enjoy Graphogame?* To reduce the possibility of students guessing the answer to the questions because of having reading difficulties, a research assistant read the questions and asked students to color in icon that matched their thoughts. For example, a smiley face represented a lot, a straight smiley face represented a little, and a sad face represented none. For those questions that asked for other answers (e.g., When would you play Graphogame?) a picture of a house and a school were given. Appendix A has a sample of the teleform used to collect responses in English and in Spanish). We administered the survey to 144 students. Students could respond to the survey either in Spanish or in English. In addition, 25% of the 144 students were randomly selected for an interview which allowed us to generalize outcomes for the entire sample in our study. An example of an open-ended question was: *How would you change the game to make it better?*

Teacher Survey**:** The purpose of the teacher survey was to learn more about the feasibility of Graphogame to be implemented in authentic settings as well as to better understand how teachers perceived student level of engagement in the game. The teacher survey was adapted from a previous survey (Ward et al., 2011). Teachers were assured anonymity in their responses both verbally and in written form. The questionnaire contained 29 questions, some included only rating items, others were open-ended. Seven questions were specifically related to teacher perception of student engagement (e.g., Graphogame impacted students. Potential ratings included *negative, no impact, or positive*). As a follow-up, teachers were asked an open-ended question such as *Please describe in detail the impact of the game for participating students.* Seventeen questions were related to the feasibility of implementing the game in an authentic classroom setting. Four additional questions were taken from a national survey on educational technology and were related to teacher perception of the use of this technology in their classrooms. Appendix XX includes the complete survey with teacher responses.

Fidelity of Implementation

We measured fidelity of implementation based on the amount of time students were engaged in playing the game excluding the amount of time they spent selecting stickers or tokens. Although we asked teachers to spend approximately 10 minutes per day for 16 weeks playing Graphogame, game time varied substantially among treatment classrooms (i.e., from 1.75 hours to 11 hours in the English-only classrooms, and from 4.2 hours to 14 hours in the Spanish-English classrooms). Amount of time playing Graphogame was weakly but significantly correlated with students' scores on the Graphogame Pseudoword (r = .26, p < .01) reading and Letter Sound subtests (r = .35, p < .01).

Results

Tables 2 and 3 in Appendix A2 present the means, standard deviations, and sample sizes for the reading outcomes of interest by language group and condition (i.e., treatment or comparison). We decided to report results by language group because we wanted to examine whether Graphogame had a differential effect in English on English-only students versus bilingual students given that the bilingual group received less English reading instruction than the English-only group (Baker, Park, & Baker, 2013). Based on the means and standard deviations in English and in Spanish across conditions, we can see that students in both conditions appeared to have made substantial growth in letter-sound recognition, pseudoword reading, and oral reading fluency in English and in Spanish. At posttest, English-only students in the treatment group had higher scores on DIBELS NWF and on ORF. On the other hand, bilingual students in the comparison group had higher scores on all DIBELS and IDEL measures. These differential scores were more apparent when examining the general outcome measures (i.e., DIBELS and IDEL) compared to the Graphogame measures where the difference between pretest and posttest scores was only 1-2 points.

Main Effects

To answer our first two primary research questions, we conducted Analyses of Covariance (ANCOVAs) with pretest scores as a covariate and condition (e.g., treatment or comparison) as a between-subjects factor. We analyzed the data for each outcome separately: DIBELS NWF, DIBELS ORF, IDEL FPS, IDEL FLO, Graphogame Letter Sound Evaluation (English and Spanish), Graphogame Phonology (Spanish), Graphogame Pseudoword Reading (English and Spanish), and Graphogame Word Recognition (English).

Result<u>s of the ANCOVAs for the English-only group: </u>ANCOVA analyses indicate that there were no statistically significant differences in the posttest performance of students receiving English-only instruction in the treatment group compared to the comparison group after controlling for pretest differences.

Results of the ANCOVAs for the bilingual group**:** Similarly, for students receiving bilingual instruction, results indicated no statistically significant differences in the posttest performance of students in the treatment versus the comparison condition after taking pretest scores into account with one exception. The effect of condition was statistically significant for the English Graphogame Letter Sound Evaluation task, F (1, 74) = 4.36, p = 0.04, with higher adjusted means for students in the treatment group (M_{Adj} = 17.38) compared to the comparison group (M_{Adj}=15.83). ANCOVA results also indicated that the Graphogame intervention explained only 5.6% of the variance in the Graphogame English Letter Sound scores for students receiving bilingual instruction.

Moderation Effects

We also tested a number of moderators to determine whether the Graphogame intervention was differentially effective depending on student decoding skills at pretest and amount of time playing the game. Tables 4 and 5 in Appendix A 2 present the results of the interactions between our outcome measures and our moderating variables. Below we describe the nature of these interactions.

<u>Moderating Effects of Risk-Status at Pretest on Outcomes for English-Only Students: </u>We grouped students receiving English-only instruction by risk status within conditions using the published benchmark goals for the DIBELS 6th edition measures. These goals indicate that students who earn a score between 0 and 29 on NWF are considered to be at-risk for later reading difficulties, students who earn a score between 30 and 49 on NWF are considered to be at some-risk for later reading difficulties, and students who earn a score of 50 and above are considered to be at low-risk for later reading difficulties. We present the results for all of the ANCOVAs examining the effect of condition and NWF risk status for all English reading measures in Table 4. For students receiving English only instruction, we found a significant interaction effect of risk status by condition (F (2, 165) = 3.47, p = .03, η^2 = .04) favoring students at low risk on NWF at pretest in the treatment group. In other words, level of risk on NWF at pretest moderated the effects of the Graphogame intervention for students receiving English only instruction at posttest. Students in the treatment group who were at low risk on NWF at pretest scored 17 points higher on NWF at posttest (M_{Adj} = 68.60) compared to students at low risk in the comparison group (M_{Adj} = 51.06).

<u>Moderating Effects of Risk-Status at Pretest on Outcomes for Bilingual Students:</u> Risk status for the bilingual group was determined using students' IDEL FPS score at pretest while for the English measures risk status was determined using students' DIBELS NWF score at pretest. For students

receiving bilingual instruction, results indicated that risk status on NWF at pretest moderated the effects of the Graphogame Spanish intervention for ORF scores at posttest after controlling for ORF pretest scores, F (2,76) = 3.36, p = .04, η^2 = .08. This moderating effect, however, appeared to favor students in the comparison group, meaning that the interaction between risk status and condition benefitted students in the comparison group. Scores for students at-risk in the comparison group earned scores approximately 7 points higher than students in the treatment group (M_{Adj}= 44.70 and 37.23, respectively) while some-risk students in the comparison group had scores approximately 10 points higher than their peers in the treatment group (M_{Adj} = 45.26 and M_{Adj}=35.80 respectively). Scores for students at low risk were similar in both, comparison and treatment groups (M_{Adj} = 45.62 and M_{Adj} = 44.99 respectively).

Results also indicated that, for students receiving bilingual instruction, risk status on FPS at pretest moderated the effects of the Graphogame Spanish intervention for students' scores on the Graphogame Spanish Phonology (F (2, 45) = 3.19, p = .05, η^2 = .12) and Spanish Pseudoword Reading (F (2, 66) = 5.75, p = .005, η^2 = .148) tasks at posttest after controlling for pretest scores. Results of the ANCOVA for the Graphogame Spanish Phonology task indicate that students categorized as being at low-risk on FPS at pretest had the highest Phonology scores at posttest (M_{Adj} = 13.90) compared to students at some risk (M_{Adj} = 12.90) and students at-risk (M_{Adj} = 10.54). Displaying a similar trend, results of the ANCOVA for the Graphogame Spanish Pseudoword Reading task indicated that students categorized as being at low-risk on FPS at pretest had the highest Spanish Pseudoword Reading scores at posttest (M_{Adj} = 26.49) compared to students at some risk (M_{Adj} = 24.31) and students at-risk (M_{Adj} = 16.39).

Moderating Effects of Spanish decoding skills on English outcomes***:*** As noted earlier, we also conducted ANCOVAs for students receiving bilingual instruction to determine whether their Spanish decoding skills at pretest moderated the effects of the Graphogame intervention on their English literacy performance at posttest. Results from both ANCOVAs – one with NWF Correct Letter Sounds (CLS) as the outcome and the second with ORF as the outcome – revealed that this was indeed the case. For NWF CLS the effect of FPS risk status was significant, F (2, 75) = 7.58, p = .001, η^2 = .168) and examination of the adjusted means reveals that students at low-risk on FPS at pretest had the highest NWF CLS scores at posttest (M_{Adj} = 80.34), followed by students at some risk (M_{Adj} = 61.73). Students at-risk on FPS at pretest had the lowest NWF CLS scores at posttest (M_{Adj} = 42.29). In addition, 16.8% of the variance in students' English decoding scores at posttest was explained by their level of risk on FPS at pretest. For ORF the effect of FPS risk status was also significant, F (2, 76) = 6.051, p = .004, η^2 = .137, and examination of the adjusted means reveals that students at low risk on FPS at pretest had the highest ORF scores at posttest (M_{Adj} = 46.09). Interestingly, students categorized as being at-risk on FPS at pretest had higher ORF scores at posttest than did their peers at some risk (M_{Adj} = 38.80 and M_{Adj} = 36.46, respectively). Students' level of risk on FPS at pretest accounted for 13.7% of the variance observed in their English oral reading fluency scores at posttest.

Moderating Effects of Playing Time: The amount of time students in the treatment condition played Graphogame varied widely, as evidenced by the descriptors for amount of playing time, in minutes, by language of intervention and school. Given that the amount of time varied so widely from one school to the other and across the two languages, we were interested in whether the amount of time students spent playing Graphogame was related to their posttest performance. Correlations between English playing time and performance on English reading measures were weak ranging from r = -.04 to .17, and none were statistically significant. Similar relations were observed for students receiving bilingual instruction who played Graphogame in Spanish, with correlations ranging from *r* - .17 to .36, with only the correlation between playing time and Spanish Letter Sound Evaluation at posttest being significant (*r* = .36, *p* < .01). Given that amount of playing time was not correlated with student performance at posttest, we did not examine further the moderating effect of time on the effect of the Graphogame intervention in either English or Spanish.

Student Engagement with Graphogame

The common theme of the SAVI project is understanding and mediating the relationship between engagement in learning tasks and learning outcomes. To this end, we descriptively analyzed students' responses to the questions in the student survey. Of the students who played Graphogame in English, 93 students (86.9%) indicated they thought Graphogame helped them with reading words a lot, and 88 students (82.2%) indicated they enjoyed playing Graphogame a lot. When asked if they had a choice to play Graphogame at school, at home, or not at all, 70 students (65.4%) who played Graphogame in English indicated they would play it at school, 24 students (22.4%) indicated they would play at home, and three students (2.8%) indicated they would not play it at all. We also asked students to provide information about the amount of time they played Graphogame; 26 students (24.3%) who played in English felt they played Graphogame too much, 69 students (64.5%) felt they played just the right amount of Graphogame, and one student (0.9%) indicated that he/she felt they did not play Graphogame enough. Finally, when asked how they felt about reading after playing Graphogame, 89 students (83.2%) indicated they were more excited about reading, five students (4.7%) indicated they felt the same about reading, and only two students (1.9%) indicated they didn't like reading as much.

Feedback was similarly positive for students who received bilingual instruction and played Graphogame in Spanish. For example, 90% of students (*n* = 45) indicated they thought Graphogame helped them a lot with reading words and 92% of students (*n* = 46) indicated they enjoyed playing Graphogame a lot. Seventy-six percent of students (*n* = 38) indicated that, if they had their choice, they would play Graphogame at school while nine students (18%) indicated they would play Graphogame at home; no students indicated they would rather not play Graphogame at all. Students' ratings of the amount of time they played Graphogame was also positive overall, with 16 students (32%) indicating they felt they played Graphogame too much, 30 students (60%) indicating they felt they played just the right amount of Graphogame, and only two students (4%) indicating they felt they did not get to play Graphogame enough. Finally, 48 students (96%) indicated they felt more excited about reading after playing Graphogame.

Teachers' Perceptions of GG and Student Engagement: Overall**,** teachers impressions of GG were highly positive. Results of the teacher survey indicated that teachers thought students enjoyed playing the game and also showed more excitement about reading after being exposed to the game. One of the teachers said "My students were so excited to login and play Graphogame. Over time, I even saw that my Low-level students were excited to participate and identify letter names and sounds" (see Appendix B). 60% of teachers strongly agreed that Graphogame helped their struggling readers. 100% of teachers believed GG served as a useful tool for different forms of instruction in their language arts class. Finally, 100% of teachers expressed their willingness to continue using Graphogame in their classrooms.

Conclusions

The objectives of the Graphogame study were to investigate a) whether GG was an engaging learning game for students in the United States attending low-income schools, b) whether teachers believed it could be integrated into classroom instruction with benefits to their students, and c) whether using GG improved learning.

Our findings indicate that:

1. Students were able to use the game without difficulty and were highly engaged by it.
2. Teachers reported that they would like to use GG in their classrooms and believed that students would benefit from using it
3. Students who used GG did not increase reading skills we measured relative to students who did not use it. Subsequent analyses indicated that GG appeared to provide more benefit *for children at low-risk for reading difficulties in English and in Spanish* than children who are at-risk for reading difficulties. That is, GG did not help children at risk improve their letter-sound correspondence more than the regular reading instruction provided by the school.

Future Work: The SAVI project was envisioned and planned as a two year project. Year 1 was to focus on developing and investigating feasibility and promise of each project. In year 2, we plan *to integrate MindStars Books and Graphogame into a single reading program.* I think we should also add that the MS Books are really the follow-on of the Graphogame study. The programs, when used together, *are likely to provide a comprehensive and effective reading treatment,* covering word recognition automaticity (GG), fluent reading and comprehension (MSBs).

REFERENCES

Baker, Good, Knutson, & Watson. (2006). *Indicadores dinámicos del exito 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 exito 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, Hacioglu, 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 vygotskyian 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, & Adelgais. (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: Routeledge.

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.

Mostoww, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of project listen. In K. Forbus & P. Feltovich (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 Assessmen*. 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 desig. *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

English measures

GraphoGame US English Rime Assessment. This measure has three subtests, the letter sound assessment task, rime/pseudoword recognition and word knowledge tasks. In the letter sound assessment the student has 24 trials each including seven letters for which the task is to select the one that corresponds to the presented sound. Each trial was constructed so that the seven alternative letters could be either connected to sounds that are phonetically similar to the target sound and/or that the alternative letters are visually confusable with each other. (e.g., target *o* with distractors *u p e n m a* or the target *d* with distractors *b g p f h n*). For the rime/pseudoword assessment and the word recognition tasks students are presented with 32 trials each. Students were assessed with the same measures at pretest and posttest.

DIBELS Nonsense Word Fluency (NWF). NWF is a standardized, individually administered test of the alphabetic principle. It is a subtest of the Dynamic Indicators of Beginning Early Literacy Skills (DIBELS,(Good & Kaminski, 2002; Good, Wallin, Simmons, Kame'enui, & Kaminski, 2002). Successful performance on NWF indicates knowledge of (a) letter-sound correspondences, in which letters represent their most common sounds, and (b) how to blend letter-sounds into whole units (i.e., pseudowords). According to Good and Kaminski (2002), alternate-form reliability coefficients for NWF ranged from .67 to .87, and concurrent validity coefficients with the readiness subtests of the Woodcock-Johnson Psycho-Educational Test ranged from .35 to .55. Recent studies indicated moderate correlations (r = .56) between NWF at the end of kindergarten and the SAT-10 reading comprehension subtest at the end of first grade (Fien et al., 2008; Good, Baker, & Peyton, 2009; Good et al., 2002). In this study, we administered alternate forms of NWF at pretest and posttest.

DIBELS Oral Reading Fluency 6th Edition (ORF). ORF is a standardized, timed, individually administered test of accuracy and fluency (Good & Kaminski, 2002). Oral reading fluency is designed to (a) identify children who may need additional instructional support, and (b) monitor progress toward instructional goals. Reading passages are calibrated for each relevant grade level, and the median number of words students read correctly across three different passages is reported. Students read each passage for 1 min. Words omitted, substituted, and hesitations of more than 3-s are scored as errors. Words self-corrected within 2 s are scored as accurate. In previous studies, alternate-form reliability coefficients of different reading passages from the same level of difficulty have ranged from 0.89 to 0.94 (Good & Kaminski, 2002). In Oregon, the correlation between ORF and the Oregon Assessment of Knowledge and Skills (OAKS) reading measure at the end of third grade was reported as 0.67 (Good, Simmons, & Kame'enui, 2001). In this study, we administered three passages at pretest and three alternative passages of the same reading level at posttest.

Spanish Measures

GraphoGame US Spanish Assessment. Similar to the GraphoGame English assessment, this measure has also three tasks, letter sound knowledge, phonological awareness, and pseudoword recognition. The letter sound assessment task has 21 letter sounds. Each target is presented with all the other letters in a random fixed order. The phonological awareness task has 19 targets. The task is to select the correct picture from a set of three that corresponds to a word that includes the auditorilly presented target (i.e., either a phoneme, a syllable, or a word). The pseudoword recognition task has cut off points (less than two correct in a set of 8 discontinues the task). The length of pseudowords varied from short ones such as *ta* and *le* to longer ones such as *trens* and *frier)*.

IDEL Fluidez en las Palabras sin Sentido (FPS). FPS is a subtest of the Indicadores Dinámicos del Exito en la Lectura (Baker, Good, Knutson, & Watson, 2006). It is a standardized, individually administered test of the alphabetic principle and it is similar in structure to the DIBELS NWF in English. An important noticeable difference between NWF and FPS is that on the FPS, CV and CVCV nonsense words were used (e.g., lu, mosi), whereas on the NWF task, VC and CVC nonsense words were used (e.g., ug, lut). In a pilot study, the 3-week, alternate-form reliability of FPS in the middle of first grade was 0.76 (Baker, Good, Peyton, & Watson, 2004). The concurrent validity of FPS with the Woodcock-Muñoz Pruebas de Aprovechamiento subtest of Análisis de Palabras was 0.72 at the end of first grade (Watson & Froyd, 2007). In this study, we administered alternate forms of FPS at pretest and posttest.

IDEL Fluidez en la Lectura Oral (FLO). FLO is a standardized, timed, individually administered test of accuracy and fluency with reading connected text in Spanish. It is a subtest of IDEL (Baker, Good, Knutson, et al., 2006). Passages were written taking into account sentence length, number of high frequency words, and number of letters and syllables in words. Administration and scoring of the measure is the same as those of the DIBELS ORF measure. Alternate-form reliability of different reading passages from the same level of difficulty ranged from 0.88 to 0.94. Criterion-related validity with the Woodcock-Muñoz average score was 0.75. In this study, we administered three passages at pretest and three alternative passages of the same reading level at posttest.

Quality of Instruction

We also observed reading instruction in four control classrooms and five treatment classrooms to ensure that instruction was similar in both conditions, and to record the number of minutes teachers spent on each core reading component. Although, the quality of instruction was not a factor in the analysis of the usability and feasibility of GraphoGame, we were interested in learning more about other behaviors that could potentially moderate the effects of the GraphoGame intervention. The observation instrument we used in this study was adapted from an instrument previously used in another project (Baker et al., inPreparation) and it consisted of 3 parts. The first part included teacher and school information. The second part included 8 items addressing the content of the instruction. The third part included a checklist of teacher behaviors that have been found to be effective instructional practices when teaching beginning reading in either English or Spanish such as providing an explanation of the task, modeling the activity, student opportunities to respond in unison, student opportunities to respond individually, and error correction (Baker, Park, Baker, & Basaraba, 2012; Coyne, Kame'enui, & Carnine, 2011; Vaughn et al., 2006).

For each of the specified behaviors in part 2, the behaviors observed were rated on a 4-point scale: *consistently, sometimes, rarely,* and *never*. We determined that a 4-point scale would be sufficient for us to detect important differences among teachers and for observers to use reliably.

Data Collection

Prior to pretesting, all data collectors received a half-day training on the administration of the DIBELS and IDEL measures by expert trainers. A one-hour refresher training was provided at posttest. After training, and in the field an administrator and an inter-rater each rated and scored a student assessment independently yet in close proximity to each other. The percentage of agreement between 2 observers when allowing for 1-point discrepancy was 100% on all DIBELS and IDEL measures. Student surveys on their level of engagement was conducted whole group. Interviews were conducted one on one and student responses were written verbatim.

Data Analysis Procedure

To examine the usability and feasibility of the game, we analyzed the results of teacher and student surveys descriptively. To examine the effect of the intervention on student performance on the GraphoGame measures, and the general outcome measures (i.e., DIBELS and IDEL), we conducted analyses of covariance (ANCOVA) on end T_2 outcomes with middle T_1 (pretest) scores as covariates. Although our unit of random assignment was the classroom, and students were nested within classrooms, we analyzed the data at the student level given that we had not enough classrooms to fully power our analysis. When deemed theoretically appropriate, the models were expanded to test moderators (e.g., student decoding skills, classroom, school, and amount of GraphoGame playing time).

APPENDIX B

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-Vygotskyan 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 meditated. 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.

<u>Science is special:</u> 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 then 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 its 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 dialogs 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 the 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.

C.4

UNIVERSITY OF COLORADO: ATLAS ASSESSMENT AND RESEARCH

Results of IES CASL 2015 Replication Study

My Science Tutor: Improving Science Learning Through Tutorial Dialogs (MyST)

Comparison of Learning Outcomes following Tutoring with Human Tutors or a Virtual Tutor

Tim Weston

3/15/2015

Table of Contents

Executive Summary

The assessment for the IES project*: My Science Tutor: Improving Science Learning Through Tutorial Dialogs (MyST)* was conducted from October to May during the 2013 – 2014 school year.

Students completed pre/post FOSS assessments for these modules before and after the classroom instruction and tutoring. Students participating in the study received tutoring from MyST individually, or from the human tutors in groups for 25-minute sessions concurrent with their regular classroom instruction in FOSS.

The FOSS assessments for the five modules used in the assessment have identical pre and post versions with open-ended, short answer, multiple choice and graphing items. The NWEA tests were 10 item multiple choice tests created for BLT by the testing company to cover the same topics. We gave the tests to students before and after the use of the modules and the classroom instruction.

A subset of open-ended questions on the FOSS tests were graded by two independent raters. The interrater reliability (ICC) equaled .93. Internal reliabilities (Cronbach's Alpha) ranged from $\alpha = .82$ to $\alpha = .88$ across assessments for each module for the FOSS test, but were lower for NWEA tests with alphas between .57 and .77 .

Six-hundred-one (601) students participated in the study during 2013-2014. Of these students, 211 were assigned to the "Human" experimental condition and 390 were in the "MyST" condition. Assignment occurred at the classroom level. Thirteen schools participated with 31 teachers, 15 in the Human condition, 16 in the MyST condition. Twenty teachers taught fourth grade students and 11 taught 5th grade.

Effect sizes for pre/post gain favored the human group: for the FOSS test the effect size equaled .37, for the NWEA test the difference was .22.

Gain differed by module. For the FOSS test, the biggest differences in gain was for the *Electricity and Electromagnetism* module where the difference in gains for this module was .93 of a SD unit favoring the Human group. The MyST group gained more on the *Mixtures* module than the Human group with a difference of .83. Both module comparisons showed two obvious class outliers; when these were removed the differences lessened to .53 for *Electricity and Electromagnetism* and .53 for *Mixtures* modules.

For the NWEA test, differences mirrored those of the FOSS tests, although effects were smaller. *Electricity and Electromagnetism* favored the Human group with a .36 difference although the *Living Systems* module showed a larger effect at .80 for the Human group. *Mixtures* favored the MyST group with a .19 difference.

(Executive summary cont.)

Significance tests used in Hierarchical and Mixed models showed no significant statistical differences for the main effect for condition for both FOSS and NWEA tests.

Significant interaction effects between condition and module were observed for both FOSS and NWEA measures. This indicates that type of tutoring had different effects depending upon module.

Multiple comparisons showed significant differences for the *Electricity and Electromagnetism* module favoring the Human group and the *Mixtures* module favoring MyST, both for the FOSS tests. The *Living Systems* module showed a significant difference favoring the Human group on the NWEA test.

When the *condition* and interactive effects for *condition by module* were tested with non-hierarchical Repeated Measures and ANCOVA procedures, the main effect for condition for the NWEA test became significant. The effect for the FOSS test remained non-significant. This outcome may point to a lack of power for the hierarchical procedure.

The 2013-2014 school year assessment for the Institute of Education Sciences funded *My Science Tutor: Improving Science Learning Through Tutorial Dialogs (MyST)* examined the effects of human and virtual tutors on student test scores in science. The experimental design compared students receiving human tutoring in groups with those students using the computerized tutor (My Science Tutor "MyST") individually. During the first year assessment for the project, classrooms were randomly assigned to tutoring conditions and students were tested pre and post with two tests.

The hypothesis for the main assessment study, based on the objectives of the developers at Boulder Language Technologies, was that students would learn roughly as much from the computer as a human tutor.

1. Program description

All students in the 2013-2014 study received in-class instruction in at least one of the FOSS modules *Electricity and Electromagnetism (EE) (4th grade), Living Systems* (LS)*, Mixtures and Solutions* (MX)*, Sun, Moon and Planets* (SMP) and *Soils, Rocks and Landforms* (SRL). All teachers followed module lesson plans and used Full Option Science Systems (FOSS) materials with lessons lasting between one to three months during the school year. The FOSS curriculum is used widely throughout the United States to teach hands-on science to elementary school students.

Students participating in the study received tutoring from MyST or from the human tutors for 25-minute sessions outside of class but concurrent with their regular classroom instruction. Students in the MyST condition were tutored individually on computers and listened and responded to questions with microphone headsets and headphones for the same amount of time and sessions. MyST uses language recognition of student speech to customize questions asked of students and the sequence of animations and other visuals embedded in the modules.

Student with human tutors worked in groups of two or three. Tutors used the same animations and other graphics used on the computer. All tutoring from MyST was keyed to the FOSS content recently covered in class and followed the "Questioning the Author" (QtA) format where students are asked open-ended questions about the content and asked to explain scientific concepts. QtA is meant to allow students to explain concepts themselves instead of havinga tutor directly provide information about a lesson. Both MyST and human tutoring employed computerized animations, interactive exercises and multiple-choice questions linked to FOSS content.

2. *Measures and Scores*

The FOSS assessments for the five 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.

The NWEA tests were 10 item multiple choice tests created for BLT by the testing company to cover the same topics. These tests were meant to be more distal and generalized from the content of the modules than the FOSS tests. These were also given pre and post at the same time as the FOSS tests.

2.1 Standardization

Because module tests have different scales (see Tables 1 & 2), 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 student's score. This difference was then divided by the average standard deviation for pooled pre and post for each test. Information about each test is presented in Tables 1 and 2.

Table 1 Means, Standard Deviation, Pre/Post Average and Scale for FOSS tests.

	Mean (pre/post combined)	**Standard Deviation**	**N (all pre/post)**	**Scale**
Electricity and Electromagnetism	29.09	8.76	622	0 -49
Living Systems	29.66	7.04	271	0 -49
Mixtures and Solutions	29.68	8.10	243	0 -49
Sun, Moon and Planets	41.54	7.26	151	0 -54
Soils, Rocks and Landforms	35.17	7.02	125	0 -49

Table 2 Means, Standard Deviation, Pre/Post Average and Scale for NWEA tests

	Mean (pre/post combined)	**Standard Deviation**	**N (all pre/post)**	**Scale**
Electricity and Electromagnetism	6.33	2.41	666	0 -10
Living Systems	6.63	2.66	355	0 -10
Mixtures and Solutions	7.13	2.34	168	0 -10
Sun, Moon and Planets	6.76	2.22	84	0 -10
Soils, Rocks and Landforms	5.59	2.14	287	0 -10

2.2 Test reliability

Pairs of raters from Boulder Language Technology scored all assessments. Raters trained together with scoring rubrics for open-ended items 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 .85 to .96, with an average Intraclass Correlation Coefficient of .93. Internal reliabilities (Cronbach’s Alpha) ranged from $\alpha = .82$ to $\alpha = .88$ across assessments for each module for the FOSS test, but were lower for NWEA tests with alphas between .57 and .77 . For our analysis of outcome scores, if scores differed between raters on open-ended items, we used the average between pairs of open-ended items across pairs of raters.

Table 3 Internal Reliability Coefficients by module

Module	FOSS	NWEA
EE	.83	.70
LS	.86	.57
MX	.88	.77
SMP	.85	.71
SRL	.82	.65

We also examined the difficulty of individual module tests through the average difficulty of items within modules and the RIT difficulty score for each NWEA test.

Table 4 Average difficulty of items across modules.

Module	P NWEA	RIT of test Grade level ~ 196 for 4th , 200 for 5th	GRADE
EE	.63	200	4TH
MX	.66	202	4TH
LS	.59	203	5TH
SRP	.67	NA	5TH
SMP	.71	203	5TH

Higher "p" values indicate an easier test for the students in the group. Fourth graders taking the EE and MX module tests had comparable average difficulty, the MX test was harder for the group of fifth graders taking the LS test. The fifth graders taking the SMP tests may have had a higher ability level than the fifth graders taking an equally difficult LS test given the p value for the SMP group.

3. Sample

3.1 Study participants

Six-hundred-one (601) students participated in the study during 2013-2014. Of these students, 211 were assigned to the "Human" experimental condition and 390 were in the "MyST" condition. Thirteen schools participated with 31 teachers, 15 in the Human condition, 16 in the MyST condition. Twenty teachers taught fourth grade students and 11 taught 5th grade.

Table 5 Number of consented students in study

Condition (#)	Teacher Id	Number In Class	Number Post Foss (Consented)	Number Post Foss Matching	N Post Nwea (Consented)	N Post Nwea (Matched)
HUMAN	2	27	M	M	16	15
	5	29	10	10	17	16
	7	26	21	19	22	20
	8	20	18	16	19	17
	10	23	20	19	21	21
	11	27	24	23	25	23
	12	27	21	17	21	17
	15	25	23	22	22	19
	18	28	23	17	21	21
	19	26	21	19	25	25
	20	19	16	15	M	M
	22	26	M	M	20	18
	25	51	M	M	46	41
	28	27	14	14	25	24
	31	22	M	M	M	M
	Total	403	211	191	322	277
MYST	1	27	16	11	2	M
	3	25	23	20	25	22
	4	25	21	21	21	21
	6	25	23	22	22	20
	9	58	44	34	41	40
	13	24	22	20	21	20
	14	19	18	16	M	M
	16	29	22	19	23	22
	17	53	20	18	46	46
	21	47	22	22	44	22
	23	26	24	20	24	23
	24	21	19	19	19	19
	26	30	26	26	25	24
	27	54	49	49	M	M
	29	32	23	17	M	M
	30	29	18	14	20	19
	Total	527	390	348	383	298

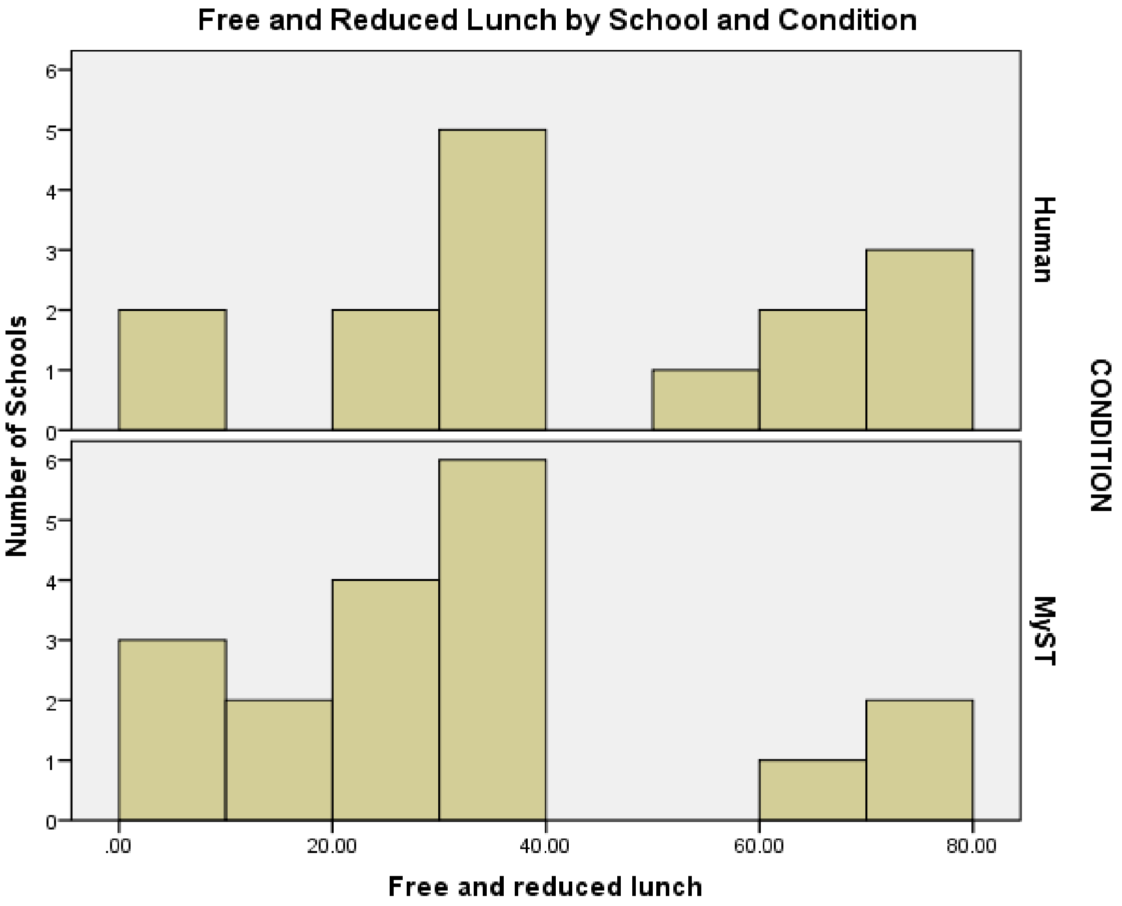
Note: M = Missing

While more students than these took the tests as part of regular instruction, for analysis, non-consented students who took the test were removed from the sample. Other tests removed were from students who did not fill out a majority of answers or provided off-task answers, and tests with grading concerns including very low reliabilities. Some teachers did not administer both tests to their students for unknown reasons. All missing data were removed by the analyst blind to tutored or control group.

3.2 Free and Reduced Lunch for participating schools.

Information about the average socio-economic status of each school came from the Colorado Department of Education website. A common proxy measure of SES for each school is the percentage of students receiving assistance from the federal free and reduced lunch program at each school; in general, the higher the percentage of students enrolled in the program, the lower the average parental income for the school. The range of percentages for participating schools is presented in figure 1. In the current data, we only saw a small correlation between post test and FRL by school at -.12.

Figure 1 Free and reduced lunch by school and condition



3.3 Gender and Native language status for tutored students.

[NEED]

4. Results

We used the standardized scores for analysis comparing average scores for each group. The two group comparisons were made for both tests using hierarchical linear models comparing pre/post test scores by *condition* and *module*, and non-hierarchical models with a Repeated Measures ANOVA and an ANCOVA.

4.1 Descriptive statistics for standardized scores.

Because not all students took both pre and post tests, we examined both matched and unmatched samples for the standardized scores for FOSS and NWEA tests.

Table 6 Standardized Pre and Post scores by condition for FOSS

		PRE	**POST**	**GAIN (Matched)**	**GAIN (Unmatched)**
Human	Mean	-.60	.79	1.46	1.39
	SD	.76	.78	.98	(1)
	Valid N	323	211	191	
MyST	Mean	-.47	.62	1.09	1.09
	SD	.75	.88	.97	(1)
	Valid N	387	390	348	
				¶ Δ = .37	Δ = .30

Table 7 Standardized pre and post scores by condition for NWEA

		PRE	**POST**	**GAIN (Matched)**	**GAIN (Unmatched)**
HUMAN	Mean	-0.56	0.55	1.10	1.11
	SD	0.97	0.84	1.06	(1)
	Valid N	341	322	277	
MYST	Mean	-0.36	0.53	0.87	0.89
	SD	0.86	0.83	0.92	(1)
	Valid N	410	333	298	
				Δ = .23	Δ = .22

While matched and unmatched samples differed by number the number of students, actual gain from pre to post was similar.[1] For the matched sample, students taking the FOSS test with human tutors on average gained .37 of a SD unit more than students using the computerized

[1] Weighted average used to find total gain for the unmatched sample. Standard deviation is assumed to equal 1 given the standardization method.

tutors. For the unmatched sample, gain was smaller at .3. For the NWEA test, the difference in gain (favoring the Human group) was .23 for the matched sample and .22 for the unmatched sample. Differences in gain were smaller when two outlier classes were removed from the sample (see below in module section for description).

Table 8 Standardized Pre, Post and Gain for FOSS test.

Comparison	Effect Size: FOSS	Effect Size: NWEA
Pre	-.13	.2
Post	-.17	.02
Gain	-.37^/.26	.22

Note: Effect size is the difference in gain between groups divided by the SD for the measure; for gain the SD is for the gain score. SD assumed to equal 1 for both measures.
^ Removal of outlier classes 11 and 25 changes this value to .26

4.2 Examination of pre test scores for class assignment by module

Because of the small numbers of classrooms using some of the modules, we examined pre scores for each class to learn how assignment of classes within each module affected differences between conditions. Because only a few classrooms used some of the modules, large differences in pre score may confound some comparisons of gain. As with many pre/post comparisons, higher gains are associated with lower pre score averages.[2] The *class* by *condition* comparison within each module for pre score are shown in figures in the appendix.

Table 9 shows pre score averages and gain for each module and condition for FOSS and NWEA scores. FOSS pre scores differed significantly for *Electricity and Electromagnetism*, and *Soils, Rocks and Landforms*. Lower pre scores were associated with higher gain for all modules except for *Living Systems*. For the NWEA measure, no significant differences were seen for pre-test, and all modules except *Mixtures* had lower pre-scores matched with higher gain.

These pre-existing differences may qualify some of the conclusions for gain in separate modules. Examination of the distribution of pre-score by class showed two possible outliers in the FOSS EE and MX modules, classes 11 and 25 (see figures 2 & 3). We retested the effects with these classes selected out in analyses below.

Table 9 Pre and Gain for FOSS and NWEA measures

[2] Correlation is not independent and so is highly inflated.

		FOSS		NWEA		
Module	Condition (#)	Pre	Gain	Pre	Gain	Difference in Gain
EE	HUMAN	-0.65*	1.73	-0.58	1.25	-.36
	MYST	-0.31	0.80	-0.32	0.89	
LS	HUMAN	-0.49	1.37	-0.63	1.47	-.80
	MYST	-0.60	0.99	-0.52	0.67	
MX	HUMAN	-0.37	0.75	-0.62	0.77	.19
	MYST	-0.64	1.58	-0.41	0.96	
SMP	HUMAN	-0.49	1.20	-0.44	0.73	
	MYST	-0.67	1.40	-0.18	M	
SRL	HUMAN	-0.78*	1.32	-0.50	1.04	
	MYST	-0.04	0.85	M	M	

Note: Differences in pre score tested with Dunnett’s T-3 using a combined categorical variable for module and condition for independent variable.

Figure 2 Pre score test by class and module for FOSS: EE

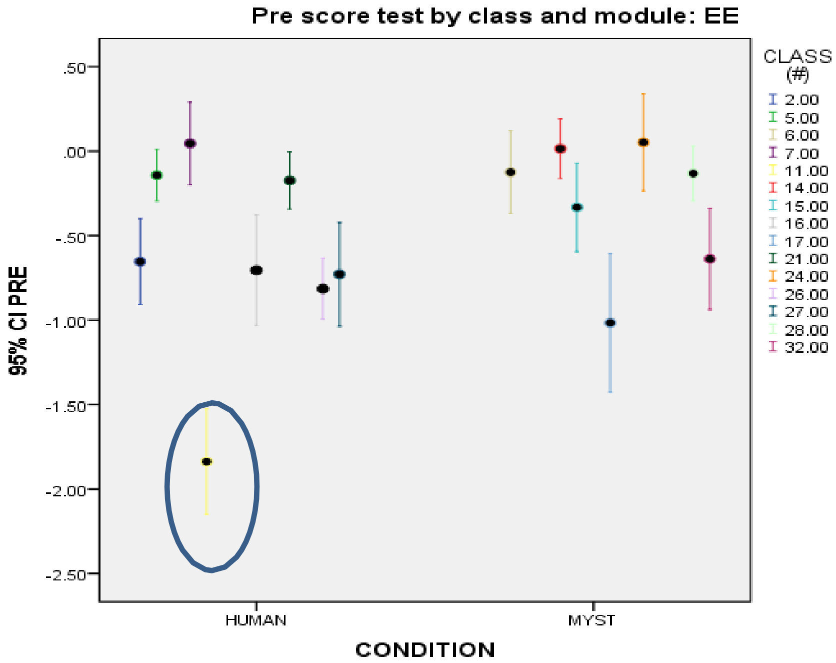
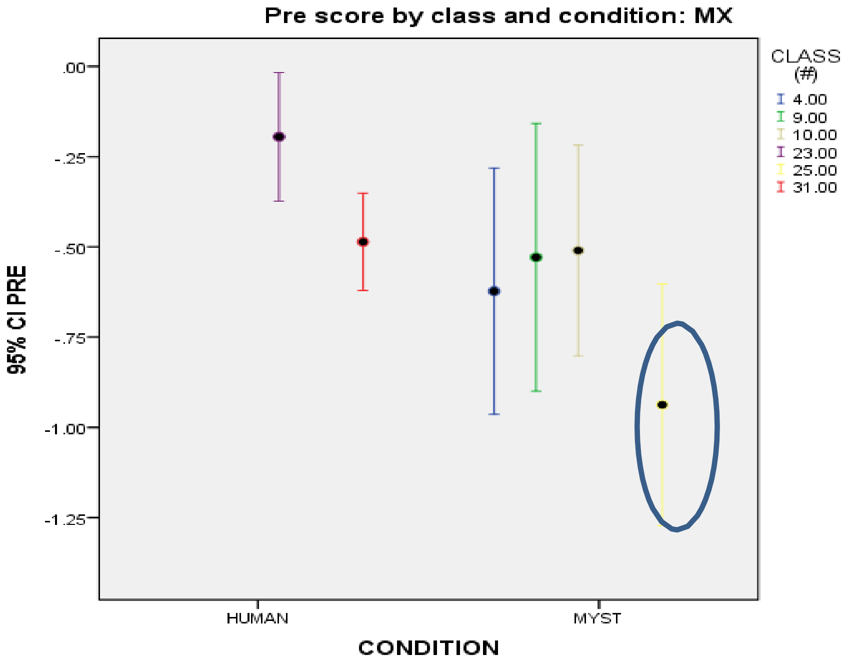


Figure 3 Pre score test by class and module for FOSS: MX



4.3 Descriptive statistics and comparison by modules

Five FOSS modules were tested with students. These included *Electricity and Electromagnetism* (EE), *Living Systems* (LS), *Mixtures and Solutions* (MX), *Sun, Moon and Planets* (SMP), and *Soils, Rocks and Landforms* (SRL). We first examined gain from pre to post for each module for the matched and unmatched samples for each condition. For the FOSS test, gain was roughly similar across modules, with larger gain for the Mixtures and Sun Moon and Planets modules. Results suggest some equivalence between the five different module tests in terms of how they functioned with their respective groups of students.

Table 10 Matched gain for FOSS and NWEA tests

	GAIN FOSS	GAIN NWEA
EE	1.15	1.06
LS	1.13	.98
MX	1.45	.90
SMP	1.35	.73
SRL	1.14	1.04

We then looked at the comparison between conditions within each module. Due to logistical considerations in data collection, the numbers of students and classrooms varied substantially between modules. The pre-post averages for each module and each condition are shown in tables 11 and 12 and figures 4 and 5.

For the FOSS test, the biggest differences in gain were for the *Electricity and Electromagnetism* module where the MyST group started off with a higher score on the pre-test than the Human group and did not score as high on the post test. The total difference in gains for this module was .93 of a SD unit. The MyST group gained more on the *Mixtures* module than the Human group with a difference of .83. Both module comparisons showed two obvious class outliers for the pre scores.

For the NWEA test, differences mirrored those of the FOSS tests, although effects were smaller. *Electricity and Electromagnetism* favored the Human group with a .36 difference although the *Living Systems* module showed a larger effect at .8. *Mixtures* favored the MyST group with a .19 difference.

Table 11 Pre, Post and gain for FOSS measure by module and condition

			PRE	POST	GAIN (Matched)	GAIN (Unmatched)	N_Classes
EE	HUMAN	Mean	-0.65	1.06	1.73^	1.71	5
		SD	0.8	0.64	1.08		
		Valid N	175	90	85		
	MYST	Mean	-0.31	0.52	0.80	0.83	7
		SD	0.7	0.92	0.8		
		Valid N	154	158	140		
LS	HUMAN	Mean	-0.49	0.9	1.37	1.39	2
		SD	0.85	0.95	0.99		
		Valid N	43	45	40		
	MYST	Mean	-0.6	0.4	0.99	1	4
		SD	0.74	0.71	0.85		
		Valid N	79	76	65		
MX	HUMAN	Mean	-0.37	0.22	0.75	0.59	1
		SD	0.36	0.45	0.49		
		Valid N	41	14	14		
	MYST	Mean	-0.64	0.92	1.58#	1.56	4
		SD	0.73	0.82	1.12		
		Valid N	82	84	74		
SMP	HUMAN	Mean	-0.49	0.42	1.2	0.91	1
		SD	0.98	0.74	0.8		
		Valid N	24	21	19		
	MYST	Mean	-0.67	0.74	1.4	1.41	2
		SD	0.84	0.59	0.76		
		Valid N	51	49	49		
SRL	HUMAN	Mean	-0.78	0.44	1.32	1.22	2
		SD	0.56	0.66	0.7		
		Valid N	40	41	33		
	MYST	Mean	-0.04	0.6	0.85	0.64	1
		SD	0.58	1.46	1.33		
		Valid N	21	23	20		

^ Removal of outlier class 11 changes this value to 1.33. Difference in gain for EE becomes -.53

# Removal of outlier class 25 changes this value to 1.28. Difference in gain for MX becomes .53

Table 12 Pre post and gain for NWEA measure by module and condition

			Pre	**Post**	**Gain**	**Gain (Unmatched)**
EE	HUMAN	Mean	-0.58	0.67	1.25	1.25
		SD	0.94	0.81	1.05	
		Valid N	160	122	113	
	MYST	Mean	-0.32	0.65	0.89	0.97
		SD	0.8	0.78	0.89	
		Valid N	186	138	132	
LS	HUMAN	Mean	-0.63	0.83	1.47	1.46
		SD	1.07	0.66	1.13	
		Valid N	44	46	40	
	MYST	Mean	-0.52	0.16	0.67	0.68
		SD	0.91	0.89	0.91	
		Valid N	68	70	64	
MX	HUMAN	Mean	-0.62	0.36	0.77	0.98
		SD	0.91	0.84	0.9	
		Valid N	48	67	42	
	MYST	Mean	-0.41	0.57	0.96	0.98
		SD	0.92	0.81	0.96	
		Valid N	108	106	102	
SMP	HUMAN	Mean	-0.44	0.3	0.73	0.74
		SD	1.06	1	1.1	
		Valid N	47	47	44	
	MYST	Mean	-0.18	0.8	.	
		SD	0.83	0.55	.	
		Valid N	48	19		
SRL	HUMAN	Mean	-0.5	0.49	1.04	0.99
		SD	0.96	0.77	0.91	
		Valid N	42	40	38	

Figure 4 Pre and post scores by module for FOSS test

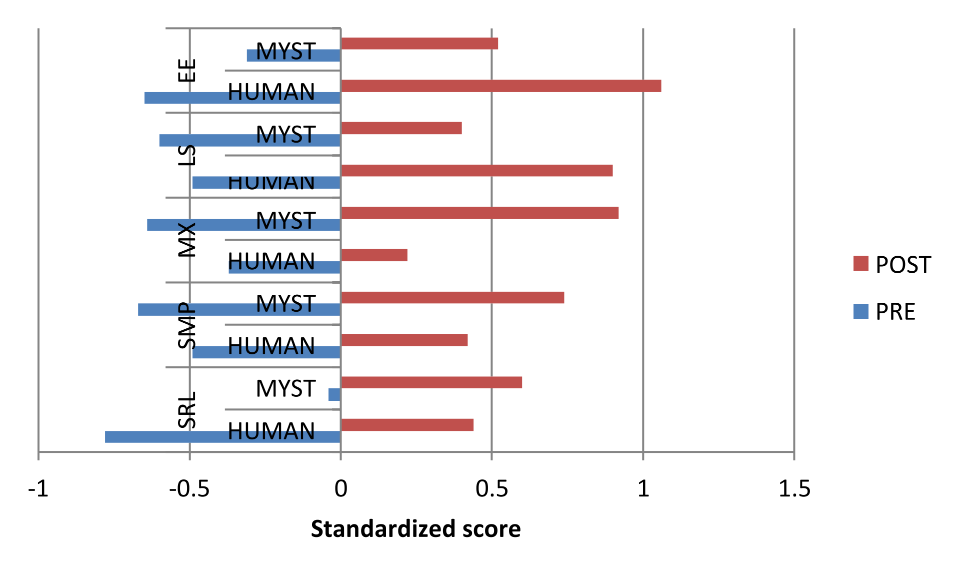
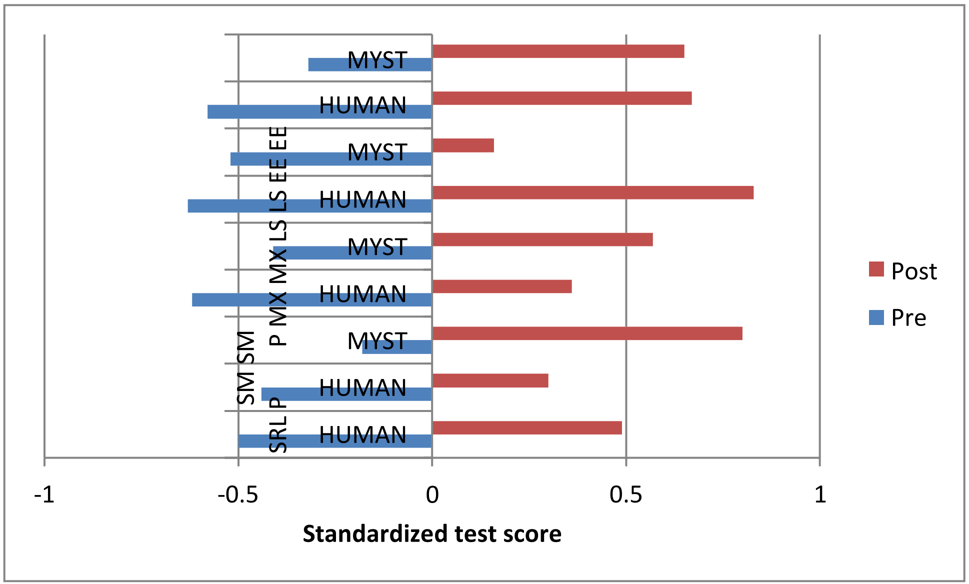


Figure 5 Pre and post scores by module for NWEA test



4.4 Hierarchical models.

The use of hierarchical models allows for partitioning of error between students and classrooms, and quantification of how much total variability is due to each level of analysis. Additionally, characteristics of each level can be tested where units were assigned. Overall is it is a fairer method of testing statistical hypotheses than traditional uni-level methods because it accounts for the effects of group membership on student scores.

The current data is structured with students within classrooms. The hierarchical model used in this study followed the form:

Level-1 Model

$$POST_{ij} = \beta_{0j} + \beta_{1j}*(PRE_{ij}) + r_{ij}$$

Level-2 Model

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(CONDITION_j) + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$

Where $POST_{ij}$ is the dependent variable post test score for each student (i) in each classroom (j), β_{0j} denotes each classroom's (j) intercept, β_{1j} the level-1 coefficient for the PRE_{ij} (prescore) covariate, and r_{ij} is the level one error. The level-2 intercept is γ_{00}, and the coefficient is γ_{01}. The variable $CONDITION_j$ is the tutoring condition (coded 0 – 1, Human = 0, MyST = 1) for each classroom with μ_{0j} the level-2 error term. Final estimates used robust standard errors.

The amount of variability due to classroom is found through the formula:

$$\rho = \tau /(\sigma^2 + \tau)$$

Where τ is the variability among group intercepts and σ^2 is student level variance.

Partitioning the variability in this manner provides information about the suitability of the data for hierarchical linear modeling. If little or no variability is due to level-2 units, the use of a multi-level model is not needed. For the FOSS data, $\rho = .23/(.46 + .23) = .33$, indicating 33% of the total variance is due to classrooms. For NWEA, $\rho= .15.$, indicating 15% of total variance is due to classrooms. In this case, both tests have substantial classroom variability.

For FOSS, the basic comparison for two groups using the hierarchical model showed no significant differences between tutored groups with a t-ratio of -1.2, *df* 27,509 and p = .23. The coefficient for the condition effect equaled γ_{01} = -.20, favoring the Human group. For NWEA, there was also no significant difference at t = -.85, p = .39 with a coefficient of γ_{01} =-.11, also favoring the human group.

Table 13 Estimation of fixed effects for FOSS measure

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	0.99	0.11	8.6	27	<0.001
CONDITION, γ_{01}	**-0.20**	**0.17**	**-1.2**	**27**	**0.234**
For PRE slope, β_1					
INTRCPT2, γ_{10}	0.30	0.05	5.6	509	<0.001

Table 14 Estimation of fixed effects for NWEA measure

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
For INTRCPT1, β_0					
INTRCPT2, γ_{00}	0.74	0.07	10.47	25	<0.001
CONDITION, γ_{01}	**-0.11**	**0.12**	**-0.859**	**25**	**0.398**
For PRE slope, β_1					
INTRCPT2, γ_{10}	0.31	0.04	7.0	547	<0.001

The main effect for *condition* with a pre-test covariate supports the primary hypothesis that the post test scores are roughly similar between Human and MyST groups.

4.5 Mixed model hierarchical comparison

Using a mixed generalized linear model we extended the previous model in an hierarchical context with the *post-test* dependent variable, a *pre-test* covariate and *condition* and *module* factors. The use of mixed models allows us to test main and interaction factorial effects while modeling the hierarchical nature of the data. [3]

The results for the FOSS measure showed a non-significant main effect for *condition* (Coefficient = -.07, p = .35); interactive effects were significant favoring the MyST group for modules *Mixtures* (Coefficient = .65, p = .004) and *Sun, Moon and Planets* (Coefficient = .16. p = .041). The overall interaction effect for *condition* by *module* was significant at F = 5.4, *df* 4, 564, p < .0001.

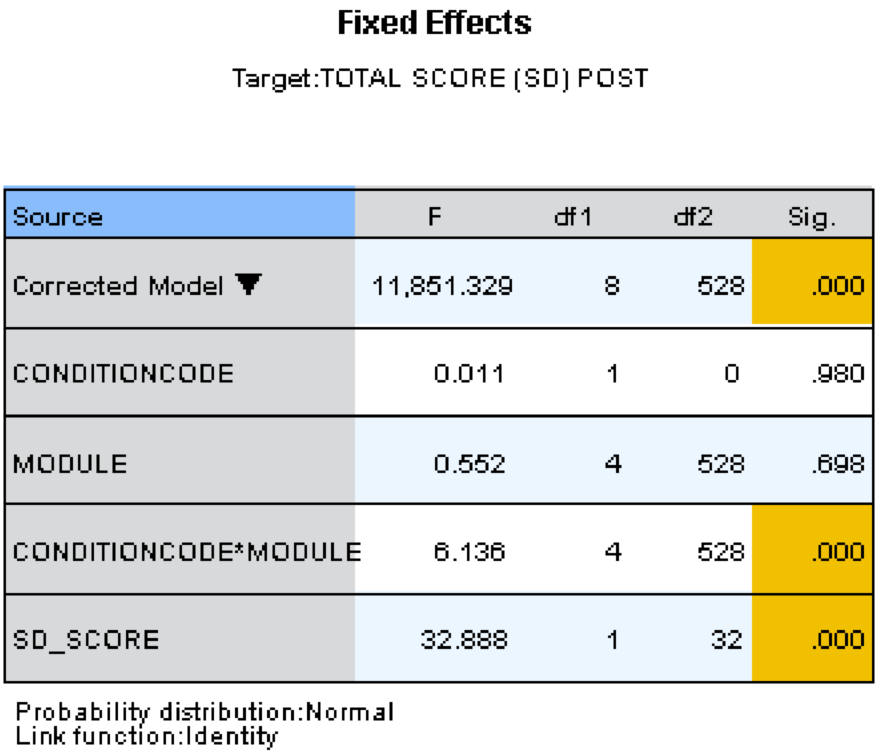
[3] In SPSS generalized linear mixed models we used the most conservative estimation procedures using the Satterthwaite approximation and robust estimation.

For the NWEA measure, the *condition* main effect was non-significant at F = 2.8, p = .12. The test for the condition coefficient -.25 and was not significant at p = .22 . The *Module* main effect was significant at F = 6.6, p < .001, and the *module* by *condition* effect was also significant at p = .014. The coefficients for the interactive effect for *Living Systems* by *condition* was also significant with coefficient = 1.03 , p = .012.

Results from the mixed models support the hypothesis that main effect for *condition* is non-significant. However, both tests show rather large interactive effects for *condition* by *module* effects. The interactions suggest that different modules have different effects on each tutoring group. Tutoring with a human seems to have more of an effect for the *Electricity and Electromagnetism and Living Systems* modules, while students using the computer may benefit more from the *Mixtures* module.

Results are also confounded by the presence of class level outliers 11 and 25 (see section 4.2) and small sample sizes (for number of classes) for the 5th grade modules SMP and SRL. Retesting without outliers with the mixed model actually gave a higher interaction term (F = 12.01) without changing the significance of the main effect for condition. Removing both outliers and the 5th grade modules increased the overall effect size to .33 but did not substantially change the significance of the main effect for condition.

Tables 15 & 16 Fixed effects and coefficients for FOSS

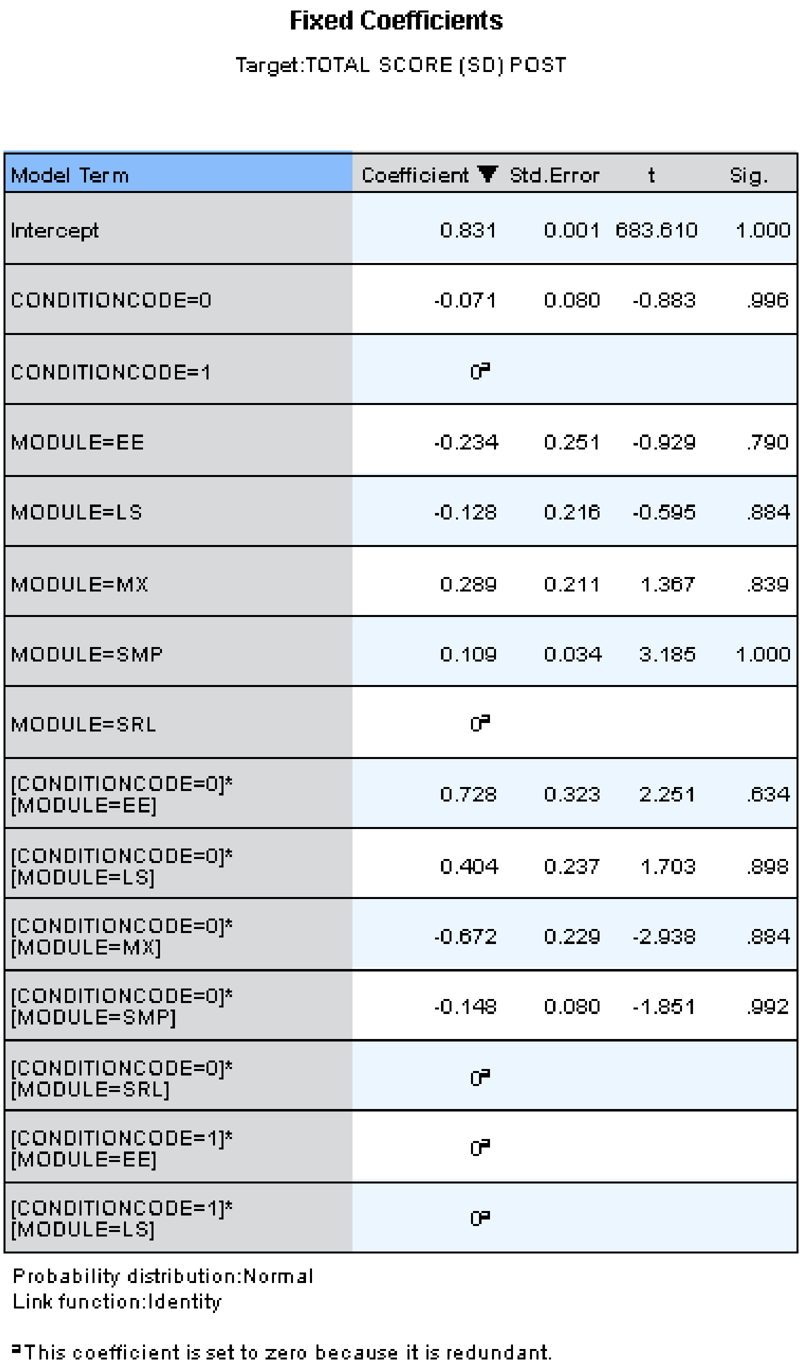


Fixed Effects

Target:TOTAL SCORE (SD) POST

Source	F	df1	df2	Sig.
Corrected Model ▼	11,851.329	8	528	.000
CONDITIONCODE	0.011	1	0	.980
MODULE	0.552	4	528	.698
CONDITIONCODE*MODULE	6.136	4	528	.000
SD_SCORE	32.888	1	32	.000

Probability distribution:Normal
Link function:Identity



Fixed Coefficients

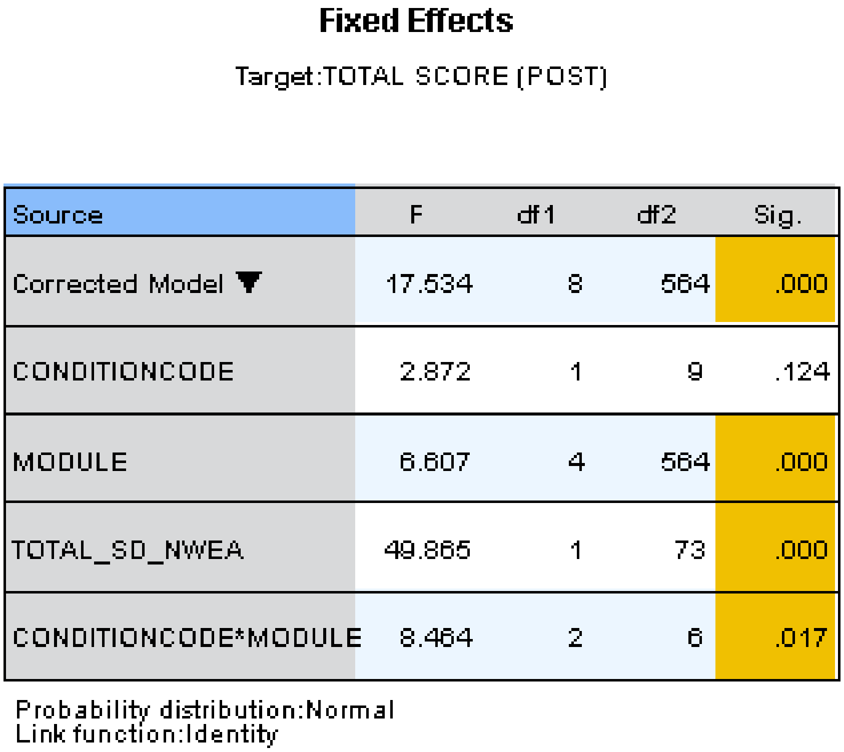
Target:TOTAL SCORE (SD) POST

Model Term	Coefficient ▼	Std.Error	t	Sig.
Intercept	0.831	0.001	683.610	1.000
CONDITIONCODE=0	-0.071	0.080	-0.883	.996
CONDITIONCODE=1	0[a]			
MODULE=EE	-0.234	0.251	-0.929	.790
MODULE=LS	-0.128	0.216	-0.595	.884
MODULE=MX	0.289	0.211	1.367	.839
MODULE=SMP	0.109	0.034	3.185	1.000
MODULE=SRL	0[a]			
[CONDITIONCODE=0]* [MODULE=EE]	0.728	0.323	2.251	.634
[CONDITIONCODE=0]* [MODULE=LS]	0.404	0.237	1.703	.898
[CONDITIONCODE=0]* [MODULE=MX]	-0.672	0.229	-2.938	.884
[CONDITIONCODE=0]* [MODULE=SMP]	-0.148	0.080	-1.851	.992
[CONDITIONCODE=0]* [MODULE=SRL]	0[a]			
[CONDITIONCODE=1]* [MODULE=EE]	0[a]			
[CONDITIONCODE=1]* [MODULE=LS]	0[a]			

Probability distribution:Normal
Link function:Identity

[a] This coefficient is set to zero because it is redundant.

Tables17 & 18 Fixed effects and coefficients for NWEA

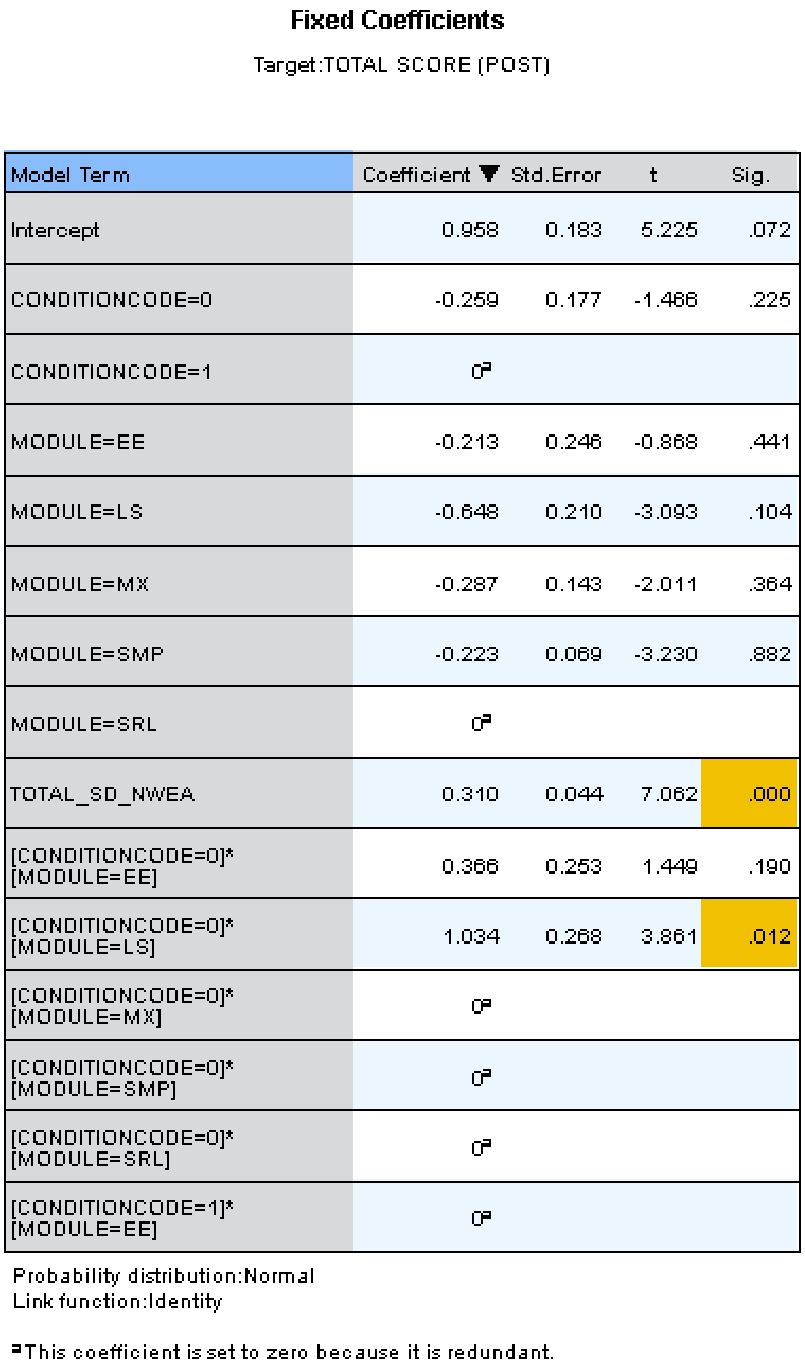


Fixed Effects

Target:TOTAL SCORE (POST)

Source	F	df1	df2	Sig.
Corrected Model ▼	17.534	8	564	.000
CONDITIONCODE	2.872	1	9	.124
MODULE	6.607	4	564	.000
TOTAL_SD_NWEA	49.865	1	73	.000
CONDITIONCODE*MODULE	8.464	2	6	.017

Probability distribution:Normal
Link function:Identity



Fixed Coefficients

Target:TOTAL SCORE (POST)

Model Term	Coefficient ▼	Std.Error	t	Sig.
Intercept	0.958	0.183	5.225	.072
CONDITIONCODE=0	-0.259	0.177	-1.466	.225
CONDITIONCODE=1	0[a]			
MODULE=EE	-0.213	0.246	-0.868	.441
MODULE=LS	-0.648	0.210	-3.093	.104
MODULE=MX	-0.287	0.143	-2.011	.364
MODULE=SMP	-0.223	0.069	-3.230	.882
MODULE=SRL	0[a]			
TOTAL_SD_NWEA	0.310	0.044	7.062	.000
[CONDITIONCODE=0]* [MODULE=EE]	0.366	0.253	1.449	.190
[CONDITIONCODE=0]* [MODULE=LS]	1.034	0.268	3.861	.012
[CONDITIONCODE=0]* [MODULE=MX]	0[a]			
[CONDITIONCODE=0]* [MODULE=SMP]	0[a]			
[CONDITIONCODE=0]* [MODULE=SRL]	0[a]			
[CONDITIONCODE=1]* [MODULE=EE]	0[a]			

Probability distribution:Normal
Link function:Identity

[a] This coefficient is set to zero because it is redundant.

4.6 Analysis of covariance (ANCOVA) & Repeated Measures ANOVA

We also tested the significance of the *module* by *condition* interaction using a (non-hierarchical) Repeated Measures ANOVA (RMANOVA) and an Analysis of Covariance (ANCOVA). We used non-hierarchical procedures, in part, as a check for the possible lack of statistical power sometimes encountered in hierarchical models. If effects are significant in level-1 analysis but non-significant in hierarchical models the differences may point to lack of power and a recommendation to add participants to future studies.

The RMANOVA tests if groups have significantly different patterns of pre/post gain across conditions and modules. For the FOSS test, the three-way interaction between *pre/post, condition* and *module* was statistically significant at F = 10.9, df 4,529, p < .0001. Table 19 presents the RMANOVA. The main effect for *pre/post* by *condition* using this model was non-significant in this model.

For the NWEA test, all main effects were statistically significant including the main effect for *pre/post* by *condition* (F = 11.09, df, 1, 667, p = .001), *pre/post* by *module* (F = 3.04, df 4, 567, p= .017) and *module* by *condition* (F = 6.7, df 4,567 p =. 001).

Table 19 Repeated measures ANOVA for FOSS measure

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Pre/Post	240.928	1	240.928	560.746	.000
Pre/Post * Module	.994	4	.249	.578	.678
Pre/Post * Condition	.943	1	.943	2.196	.139
Pre/Post * Module * Condition	18.726	4	4.681	10.896	.000
Error(factor1)	227.288	529	.430		

Table 20 Repeated measures ANOVA for NWEA measure

Tests of Within-Subjects Contrasts

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Pre/Post	188.928	1	188.928	398.318	.000
Pre/Post * Condition	5.264	1	5.264	11.099	.001
Pre/Post * Module	5.769	4	1.442	3.041	.017
Pre/Post * Module * Condition	6.697	2	3.348	7.059	.001
Error(factor1)	268.935	567	.474		

We also looked at the non-hierarchical ANOVA model with *Post-test* as dependent variable, *condition* and *module* as factors, and the *pre-test* as a covariate. For the FOSS test, this model gave similar results to the mixed model with the interactive effect for *condition* by *module* significant with F = 10.6, *df* 4, 539 and p < .0001. The main effect for *condition* in this model was also non-significant with F = .003, df 1,539, p = .954.

Table 21 ANCOVA for FOSS measure

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	56.972[a]	10	5.697	9.742	.000
Intercept	179.800	1	179.800	307.447	.000
Pre	29.030	1	29.030	49.640	.000
Condition	.002	1	.002	.003	.954
Module	2.059	4	.515	.880	.475
Condition * Module	24.951	4	6.238	10.666	.000
Error	308.782	528	.585		
Total	641.311	539			
Corrected Total	365.754	538			

a. R Squared = .156 (Adjusted R Squared = .140)

For the NWEA measure, all main effects and the interaction for condition by module were significant with values similar to the repeated measures analysis (*condition*, F = 7.1 *df*1, 575, p = .008; *module* F =4.1 *df*,4, 575; p =.003, *module* by *condition* F = 12.8, *df* 3, 575, p < .0001).

Table 22 ANCOVA for NWEA measure

Tests of Between-Subjects Effects

Dependent Variable: TOTAL SCORE (POST)

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	78.889[a]	8	9.861	17.086	.000
Intercept	145.887	1	145.887	252.780	.000
Pre	54.490	1	54.490	94.416	.000
Condition	**4.106**	**1**	**4.106**	**7.115**	**.008**
Module	9.486	4	2.372	4.109	.003
Condition * Module	14.800	2	7.400	12.822	.000
Error	326.655	566	.577		
Total	567.844	575			
Corrected Total	405.544	574			

a. R Squared = .195 (Adjusted R Squared = .183)

We then tested these effects between each condition within each module using a Dunnett –T multiple comparison using *gain* as a dependent variable. In this comparison, the *Electricity and Electromagnetism* module showed a significant difference between conditions with a mean difference of .93, significant at p < .0001; the *Mixtures* module had a significant difference favoring the MyST group with a mean difference of .79 and p = .009.

Table 23 Multiple comparisons for FOSS

MyST	Human	Mean Difference (I-J)	Std. Error	Sig.
EE	EE	-.93*	0.13	.0001
LS	LS	-0.37	0.18	0.8
MX	MX	.83*	0.18	.002**
SMP	SMP	0.46	0.32	.994
SRL	SRL	-0.19\	0.21	1

Table 22 Multiple comparisons for NWEA

MyST	Human	Mean Difference (I-J)	Std. Error	Sig.
EE	EE	0.36	0.12	0.106
LS	LS	.79*	0.21	0.009*
MX	MX	-0.18	0.16	1

Table 24 Summary of findings for FOSS and NWEA

Model	FOSS Condition	NWEA Condition	FOSS Module	NWEA Module	Condition by Module	Condition by Module
Hierarchical: *Condition, Pre-score*	Non-Significant	Non-Significant	-		-	
Hierarchical: *Condition, Pre-score, Module, Condition* by *Module*	Non-Significant	Non-significant	Non-Significant	Significant	Significant	Significant
Non-hierarchical, RMANOVA: *pre/post Condition, Module, Condition* by *Module*	Non-significant	Significant	Non-significant	Significant	Significant	Significant
Non-hierarchical: ANCOVA: *Post, pre, module, module* by *condition*	Non-significant	Significant	Non-Significant	Significant	Significant	Significant

Note: Opposing results in gray.

Table 25 Summary of estimated module effects for FOSS and NWEA

	Fixed coefficients		Gain (Group difference)		Post score difference	
	FOSS	NWEA	FOSS	NWEA	FOSS	NWEA
Electricity and Electromagnetism	-.72	-.36	-.93	.36	-0.54	-0.02
Living Systems	-.4	-1.03	-.38	-.8	-0.5	-0.67
Mixtures and Solutions	.67	16	.83	.19	0.70	0.91
Sun, Moon and Planets	.15	Na	0.2	Na	0.32	Na
Soils, Rocks and Landforms	-.21	Na	-0.47	Na	-.16	Na

Note: All values reflect MyST group minus Human group.

With class outliers removed

	FOSS
Electricity and Electromagnetism	-.53
Mixtures and Solutions	.53

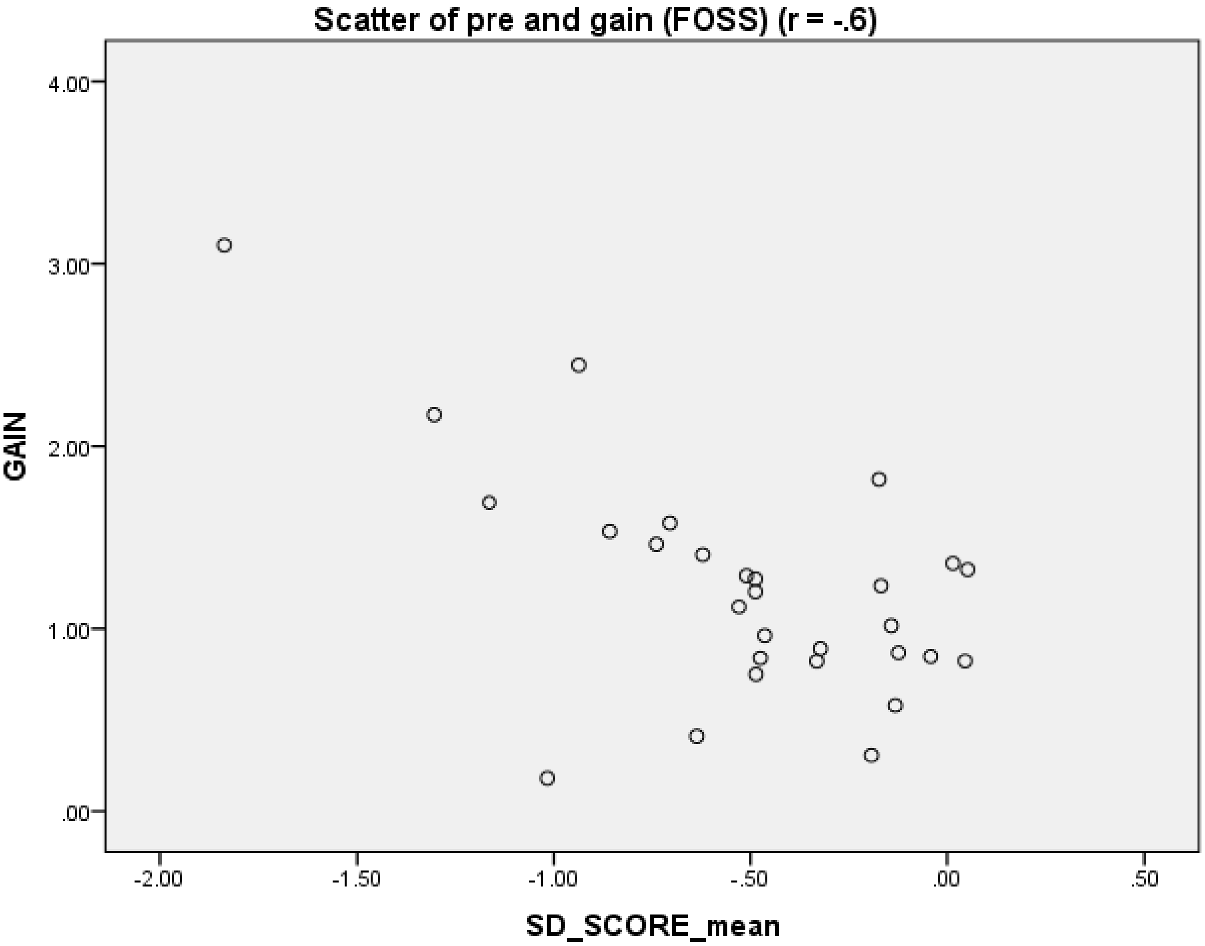
APPENDICES

5.1 Three level hierarchical model: Schools, classes, students

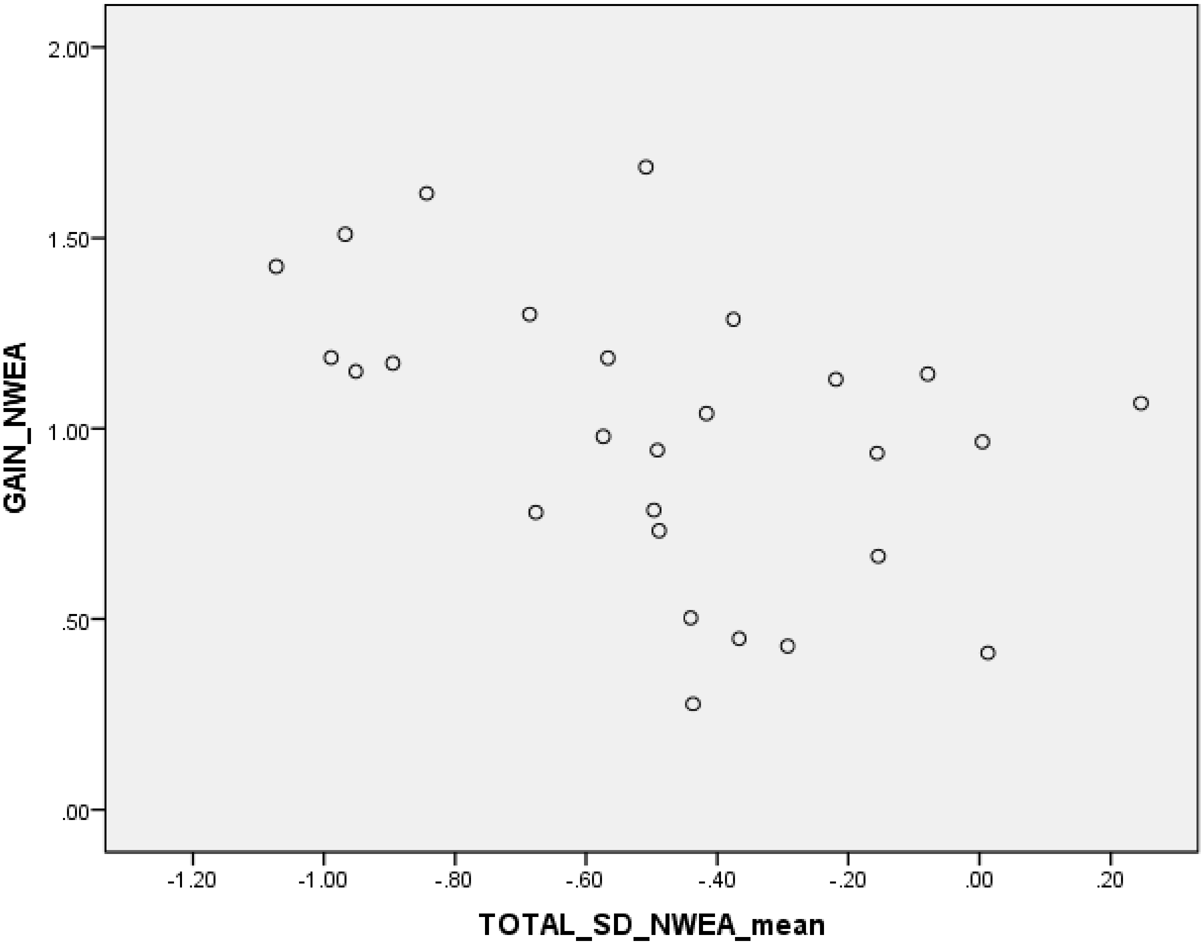
We also attempted a three level model to examine schools as a source of variance. Because the number of schools was low (n = 12), testing the effects of any school level variables (e.g. *free and reduced lunch*) was untenable. We could however estimate the amount of variability due to schools in the model for FOSS at 26% and for NWEA at 14%. Each estimate was made with the three level model and included the class by school variability in the school term.

5.2 Correlation between pre and gain

The figures present scatter of pre score and gain. Both measures show high correlations between pre and gain.



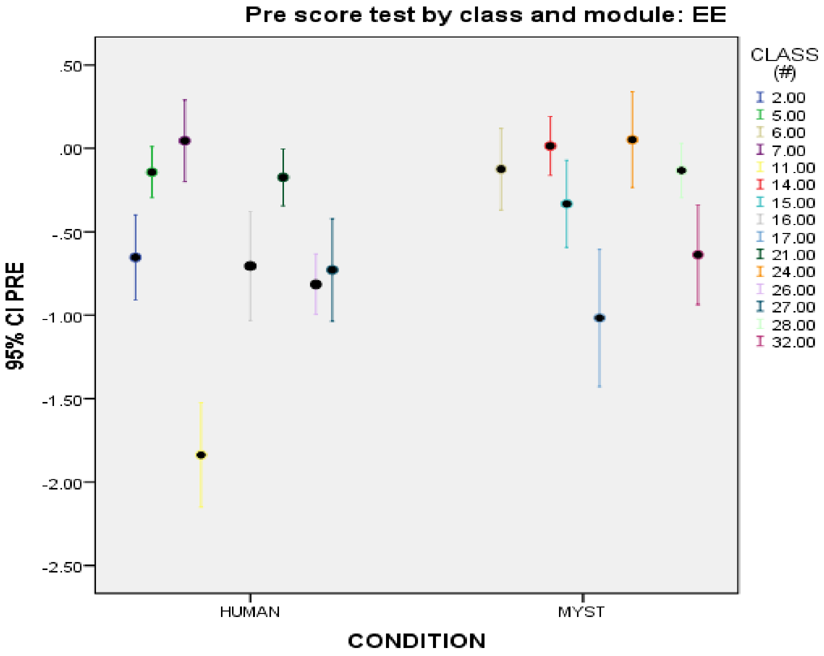
Scatter of pre and gain for class (NWEA) (r = -.45)

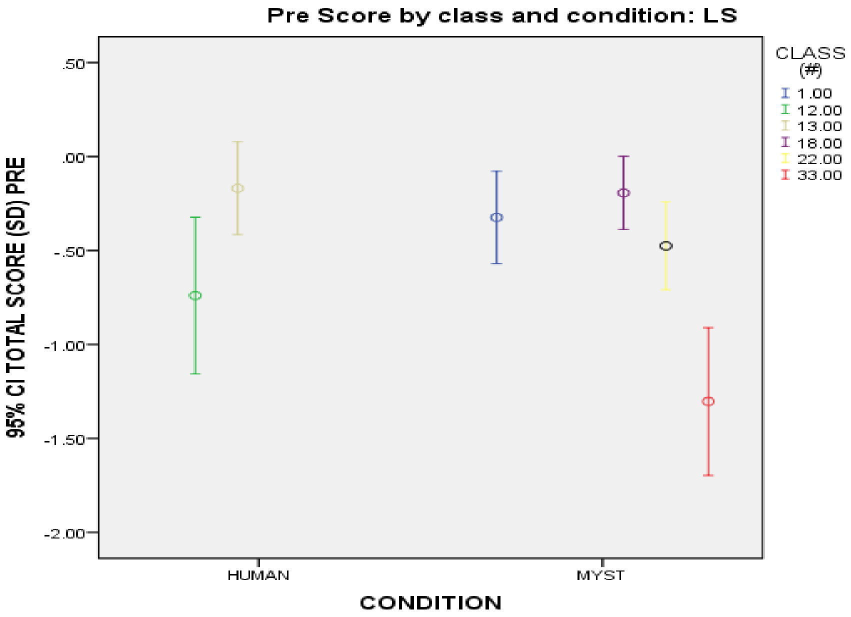


5.3 Pre-test by condition and class

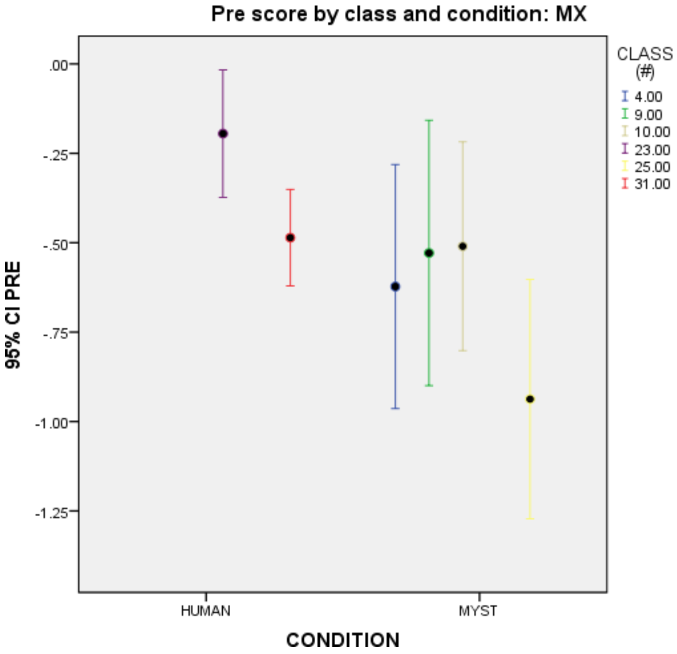
Figures show pre test averages and error for each test and each module.

(FOSS)

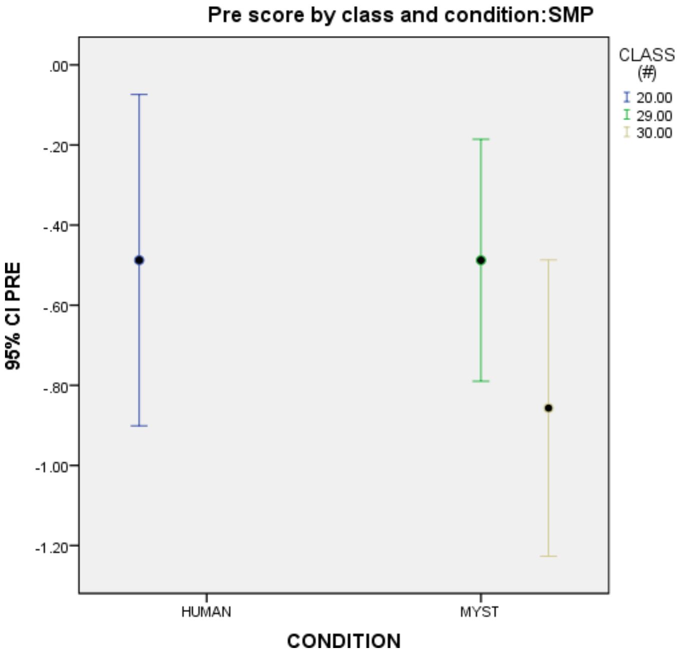


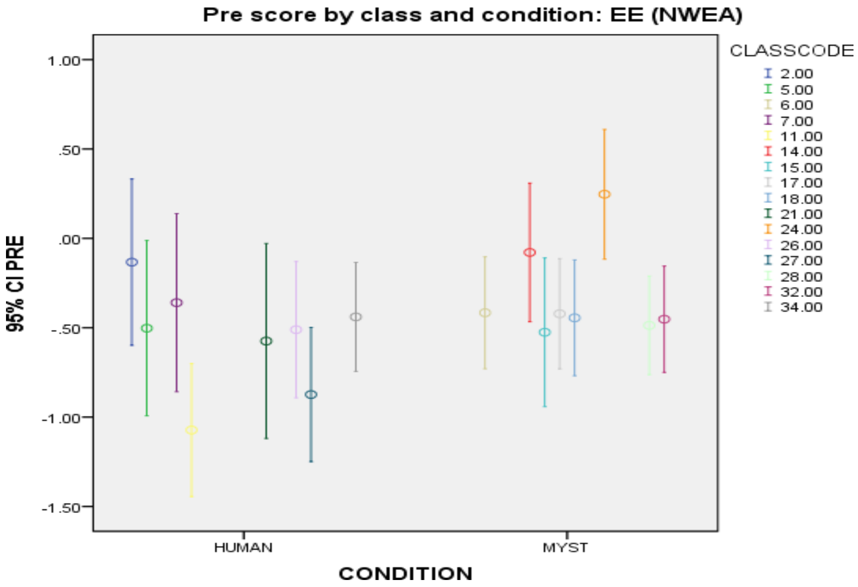


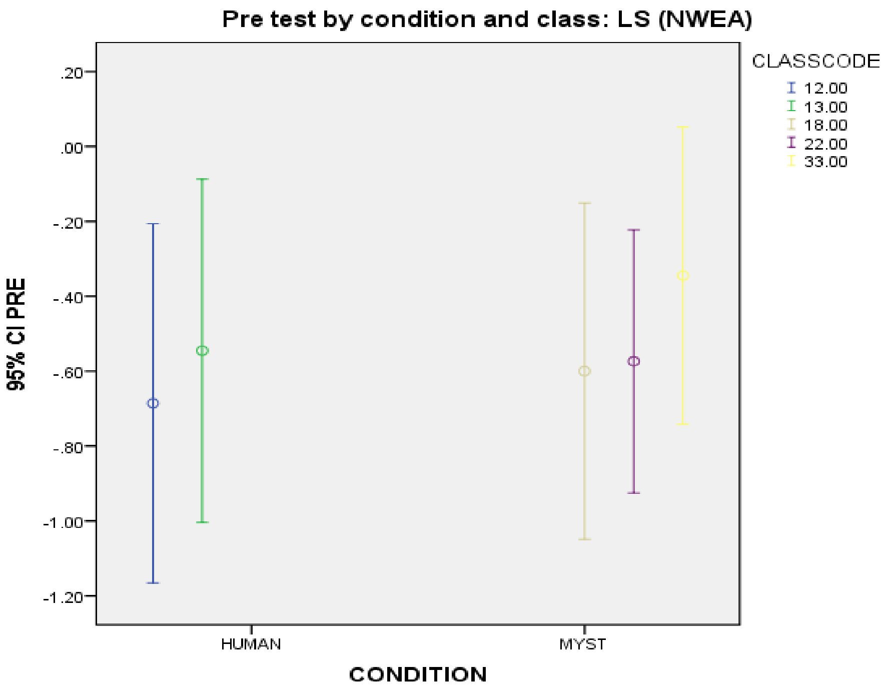
Pre Score by class and condition: LS
95% CI TOTAL SCORE (SD) PRE
.50
.00
-.50
-1.00
-1.50
-2.00
HUMAN
MYST
CONDITION
CLASS (#)
1.00
12.00
13.00
18.00
22.00
33.00

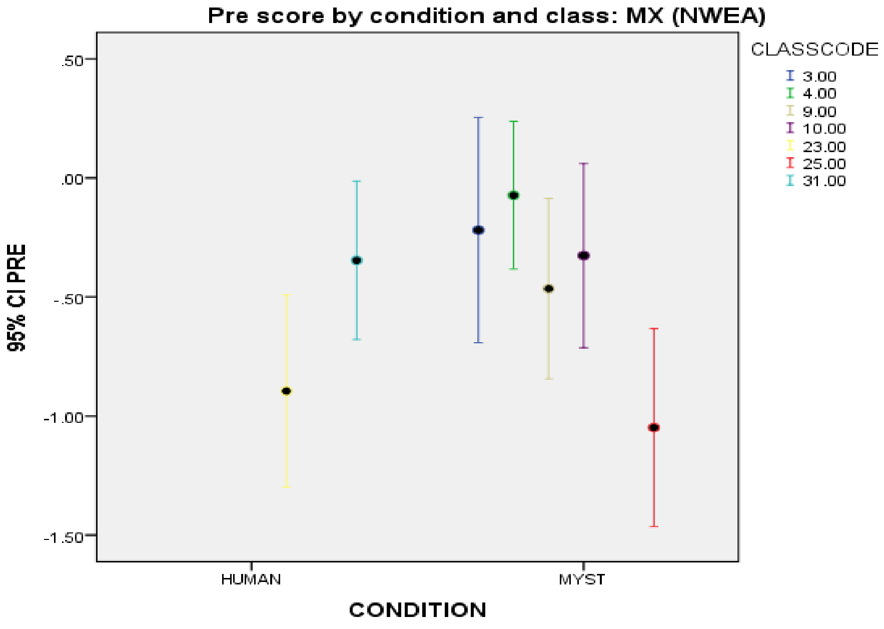


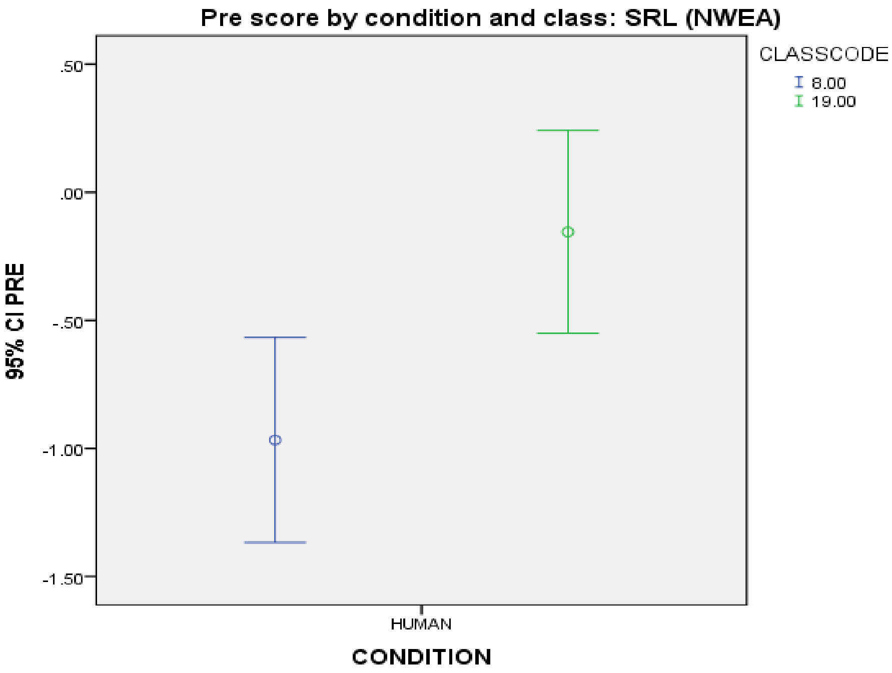
Pre score by class and condition: MX
95% CI PRE
.00
-.25
-.50
-.75
-1.00
-1.25
HUMAN
MYST
CONDITION
CLASS (#)
4.00
9.00
10.00
23.00
25.00
31.00

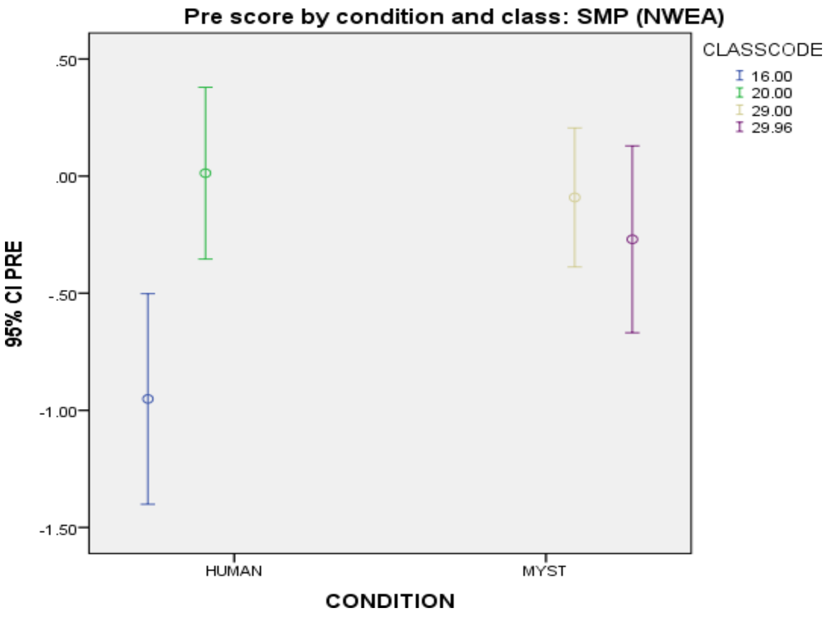

Pre score by class and condition:SMP
95% CI PRE
.00
-.20
-.40
-.60
-.80
-1.00
-1.20
HUMAN
MYST
CONDITION
CLASS (#)
20.00
29.00
30.00


Pre score by class and condition: EE (NWEA)
95% CI PRE
1.00
.50
.00
-.50
-1.00
-1.50
HUMAN
MYST
CONDITION
CLASSCODE
2.00
5.00
6.00
7.00
11.00
14.00
15.00
17.00
18.00
21.00
24.00
26.00
27.00
28.00
32.00
34.00


Pre test by condition and class: LS (NWEA)
95% CI PRE
.20
.00
-.20
-.40
-.60
-.80
-1.00
-1.20
HUMAN
MYST
CONDITION
CLASSCODE
12.00
13.00
18.00
22.00
33.00


Pre score by condition and class: MX (NWEA)
95% CI PRE
.50
.00
-.50
-1.00
-1.50
HUMAN
MYST
CONDITION
CLASSCODE
3.00
4.00
9.00
10.00
23.00
25.00
31.00


Pre score by condition and class: SRL (NWEA)
95% CI PRE
.50
.00
-.50
-1.00
-1.50
HUMAN
CONDITION
CLASSCODE
8.00
19.00


Pre score by condition and class: SMP (NWEA)
95% CI PRE
.50
.00
-.50
-1.00
-1.50
HUMAN
MYST
CONDITION
CLASSCODE
16.00
20.00
29.00
29.96

C.3

LEARNING SCIENCE THROUGH SPOKEN DIALOGS WITH A VIRTUAL TUTOR

IES CASL Innovation and Development Grant

Award R305B070008

Final Report

Project Objectives

The specific objectives of this project were:

1. Development and refinement of an intervention based on a virtual tutoring system
2. Creation of a corpus to support training and evaluation of system components
3. Analysis of the potential of the intervention based on student learning gains
4. Demonstration of the usability and feasibility of the intervention

The Intervention - MyST

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 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 think, reason and talk about science during spoken dialogs with Marni.

Using MyST, students engage in natural spoken dialogs with Marni, a lifelike 3-D computer character that is “on screen” at all times. 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 kit-based science investigations that are part of FOSS (Full Option Science System, FossWeb, 2010). In addition to the science kits that support an average of 16 hour-long 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 after each eight to ten week 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.

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 [Beck et al. 1996; McKeown and Beck 1999; McKeown et al. 1999]. QtA is a mature, scientifically-based and effective program used by hundred 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 on, 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 and 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.

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?”

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 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.

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, or an electromagnet picking up or dropping iron objects (Figure 4). 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 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. The student's computer shows a full screen window that contains the virtual tutor Marni, a display area for presenting information 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. Some displays 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 is expected to know based on the content, instructional activities and learning objectives of each classroom science investigation in each FOSS module. 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 and Aist 1999], 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.

Tutorial Development

The tutorial development process began with collection and annotation of dialogs between human tutors and students. These data were used to train a speech recognizer to recognize the words that students produce during tutoring sessions, to develop natural language processing system to interpret spoken utterances, and 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 project. 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 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.

WOZ system and collection

Our development strategy was to model spoken dialogs from 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. For all WOZ data collected, sessions were monitored by project tutors who served as the Wizards. The data from WOZ sessions was 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 offices) 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:

- 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. 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 are having difficulty expressing their knowledge of science.

System Development

MyST incorporates a number of technologies including speech recognition, dialog management, character animation, speech output, video output and presentation of flash animations. 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 visual and 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.

Data were collected in three basic conditions:

1. Human Tutor – A human tutor conducts a tutorial with a student. The human tutor has access to the visuals and other supplementary materials, but the tutor talks directly

with the student. Student speech is recorded and transcribed. 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 earlier. The WOZ system is used to gather data that more closely models the target interaction, students with virtual tutors. 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 being used in the assessment of the MyST system in schools.

Table 1 shows the data collected for each module.

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.

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.

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 was 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 6235 words. The Word Error Rate (WER) for the recognizer on the Evaluation set is shown in Table 2 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%. For the individual modules, the WER for ME and MS were very similar, while the WER for VB was substantially higher. Using a global LM, the perplexity of each module was: 56 for ME, 63 for MS and 74 for VB. Even though the ME data had a lower perplexity than the MS, the WERs are similar. VB had a substantially higher perplexity and WER. The higher perplexity of the VB data can be attributed both to less training data and to the topic of the module. The ME and MS modules are about concrete topics with which students are generally familiar. Variables introduces more abstract ideas like dependent and independent variables and graphing data. Students generally have a more difficult time with this topic, even with human tutors.

The baseline results reported above 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 and Woodland 1995] and Vocal Track Length Normalization [Lee and 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 format 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 covariances 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 2, 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. This generally would be the expected effect of recognizing more content words correctly.

Assessing Potential

During the 2010-2011 school year, we evaluated the MyST 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 (non treatment) conditions.

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.

<u>Results:</u> Direct comparisons of residual gain for the randomly assigned groups (MyST and Human Tutored) showed no significant differences between groups with $t = -1.14$, d$f = 150$, $p = 0.25$. The effect size, however, favored the human-tutored group. 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 $F = 26.2$, d$f = (2, 1164)$, $p < 0.01$. Predictably, 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). 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 ($t = 2.5$, d$f = 60$, $p = 0.014$), as did the human-tutored group when compared with controls ($t = 3$, d$f = 60$, $p < 0.01$). 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.

Demonstrating Feasibility

The MyST tutoring treatment group in the assessment study represents the proposed intervention procedure and was implemented in the intended delivery setting. 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. 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 Figure 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 also had positive things to say about MyST and its benefits to their students, even though students left the classroom when they were tutored while the teacher stayed in the classroom. Teachers reported that MyST had a positive impact on their students, that they would use the program in the future and would recommend it to other students. Interestingly, teachers indicated that, if given the choice, they would have all of their students use MyST, rather than just struggling students. Teachers were asked for feedback to help assess the feasibility of an intervention using the system and their perceptions of the impact of the system. 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. 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.

Final Report Summary

Utilizing your evaluation results, draw conclusions about the success of the project and its impact.

All project objectives were accomplished:

1. Development and refinement of an intervention based on a virtual tutoring system
2. Creation of a corpus to support training and evaluation of system components
3. Analysis of the potential of the intervention based on student learning gains
4. Demonstration of the usability and feasibility of the intervention

The MyST tutoring treatment group in the assessment study represents the proposed intervention procedure and was implemented in the intended delivery setting. The effect size of 0.53 (compared to control) would be classified as moderately strong.The student surveys indicate the usability of the system and teacher feedback speaks to its feasibility. Both students and teachers expressed overwhelmingly positive attitudes toward the system.

Describe any unanticipated outcomes or benefits from your project and any barriers you may have encountered.

What would you recommend as advice to other educators that are interested in your project?

How did your original ideas change as a result of conducting this project?

Plans for continuing the project and/or disseminating the project results

Our initial evaluation of MyST demonstrates both the feasibility and potential of the system, with positive experiences by students who used the program, positive reports by teachers on the impact of the program on their students, and a moderately strong effect (.53 Cohen d) for students who were tutored by MyST relative to students who did not receive tutoring. Based on these results, we have submitted a proposal to IES for a Goal 3 Replication and Efficacy study to demonstrate the effectiveness of MyST. The goal of the proposed study is to replicate and extend our current results with a new population of students, and thus provide good evidence for the efficacy of the program. We also plan to investigate and document possible limitations on use of the system.

Additional objectives are to demonstrate that MyST can be deployed in new schools, to demonstrate that MyST's ability to support spoken tutorial dialogs is robust to variations among students' speech patterns (vocabulary, dialects, accents) and acoustic environments (classrooms, resource rooms, computer labs), and finally, to demonstrate that MyST's tutorial dialogs can be reconfigured to align to changes in schools' science curricula in response to new programs or standards.

Figures

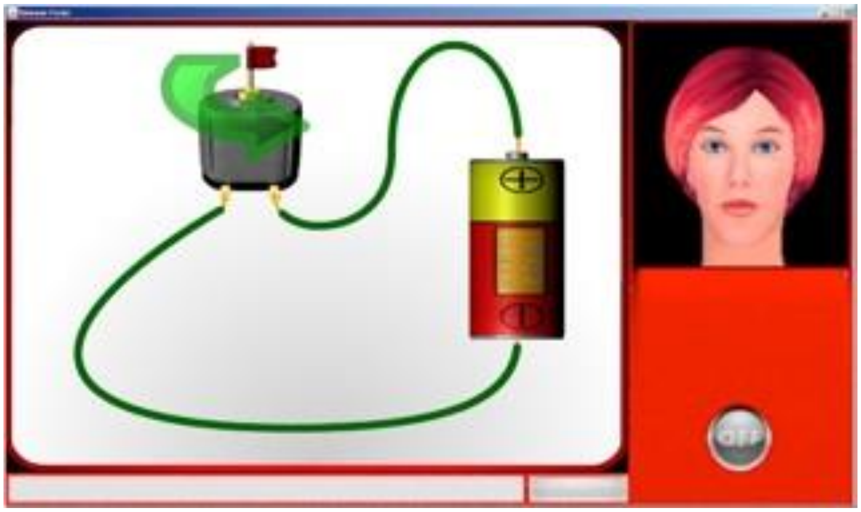


Figure 1 – Virtual Tutor Screen

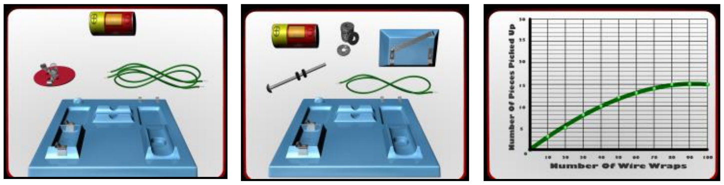
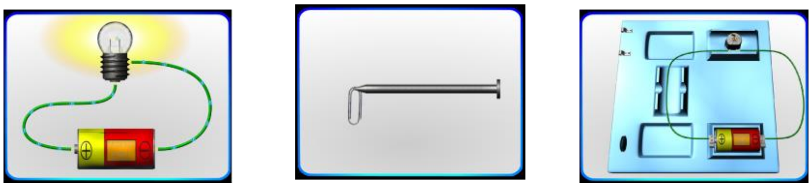


Figure 2: Example Static Illustrations

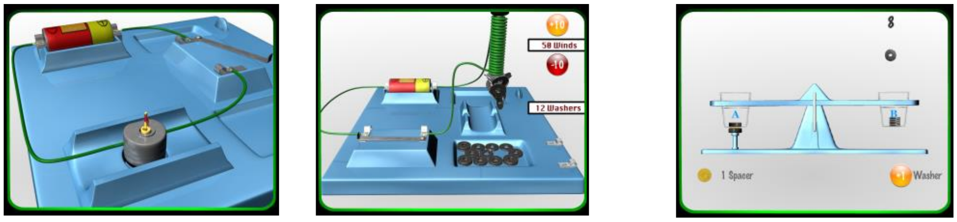


a) Electricity flowing from negative to positive terminals

b) Nail attracts clip after being rubbed by magnet

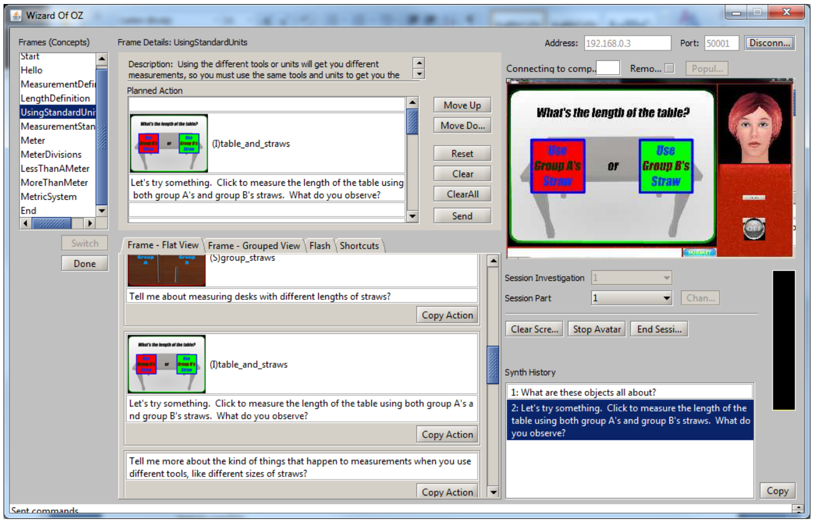
c) Motor direction depends on battery

Figure 3 – Example Animations



,

Figure 4 – Examples of Interactive Animations



Frame Details: UsingStandardUnits

Description: Using the different tools or units will get you different measurements, so you must use the same tools and units to get you the

(I)table_and_straws

Let's try something. Click to measure the length of the table using both group A's and group B's straws. What do you observe?

(S)group_straws

Tell me about measuring desks with different lengths of straws?

(I)table_and_straws

Let's try something. Click to measure the length of the table using both group A's a nd group B's straws. What do you observe?

Tell me more about the kind of things that happen to measurements when you use different tools, like different sizes of straws?

Address: 192.168.0.3 Port: 50001

Connecting to comp.. Remo... □

Session Investigation 1

Session Part 1

1: What are these objects all about?

2: Let's try something. Click to measure the length of the table using both group A's and group B's straws. What do you observe?

Figure 5 - Wizard screen

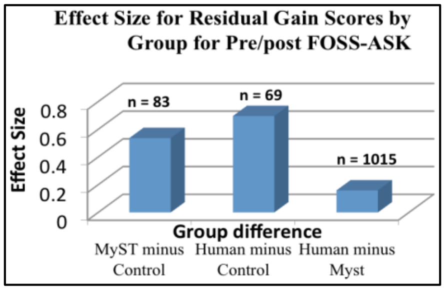


Figure 6 – Residual Gain

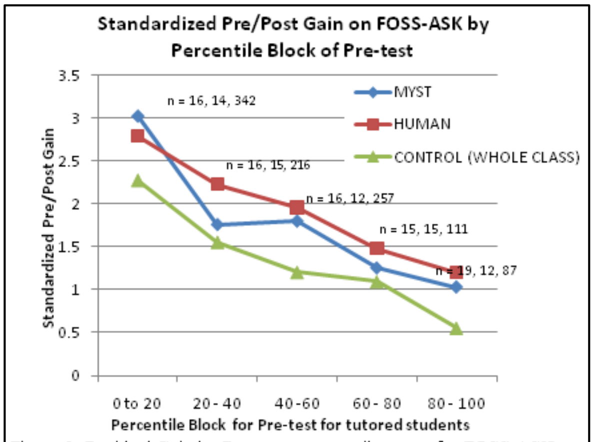


Figure 2. Residual Gain by Pre-score percentile group for FOSS ASK.

Figure 7 – Residual Gain by Pre-score

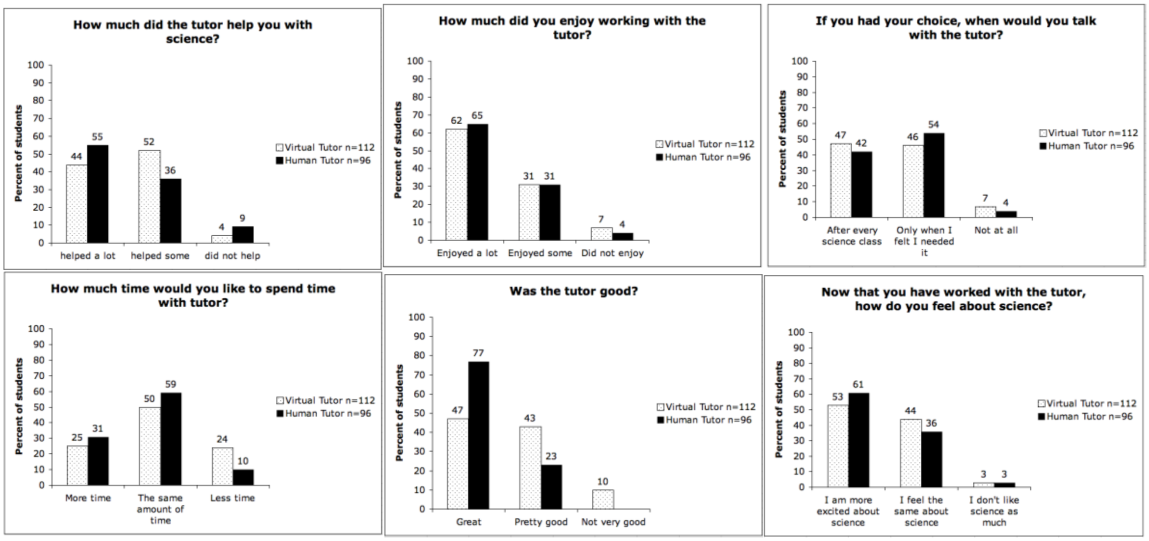


Figure 8 – Student Survey Results

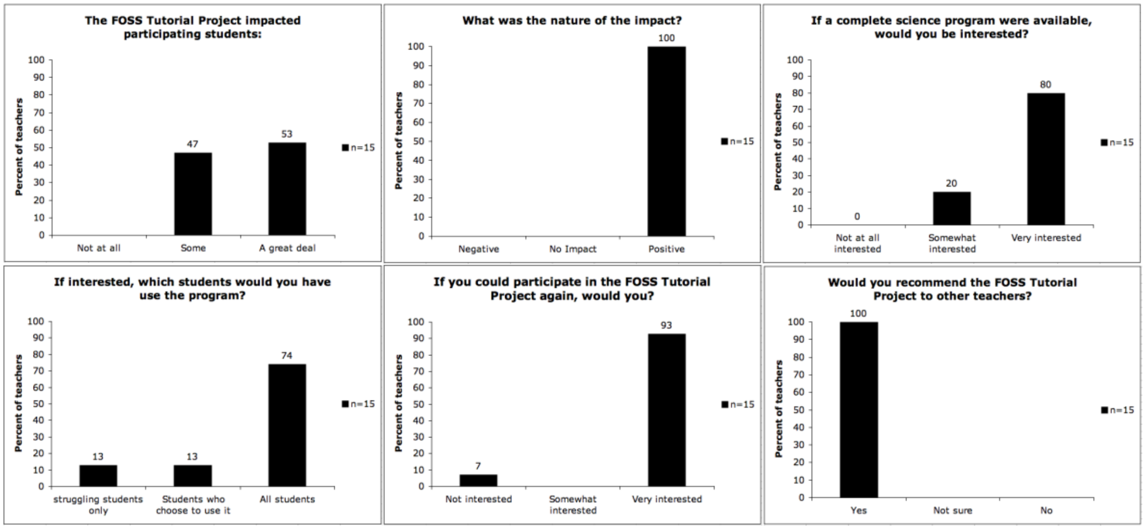


Figure 9 – Teacher Survey Results

Table 2 – 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**