One-on-one and Small Group Conversations with an Intelligent Virtual Science Tutor

Author names and affiliations:

Ronald Cole[a]
Wayne Ward[a]
Cindy Buchenroth-Martin[a]
Samantha Messier[b]
Liam Devine[a]
Timothy Weston[a]

Corresponding author:

Ronald Cole[a]

[a] Boulder Learning, Inc.
2960 Center Green Court, Suite 200
Boulder, CO 80301-5406
Phone: 303-440-3889
Fax: 303-440-3356
Email: rcole@boulderlearning.com

[b] Boulder Valley School District

DO NOT CITE Submitted for Publication

Abstract

In this study we investigated students' conversations with a virtual science tutor, either individually or in small groups. Students were presented with narrated multimedia science problems and explanations, followed by question-answer dialogs with the virtual tutor. Students who received one-on-one and small group tutoring received the same set of multimedia presentations questions posed by the virtual tutor. Students in the small group condition discussed their answer before one of the students responded to the tutor. We asked if students receiving computerized tutoring in groups would achieve learning gains equivalent to students receiving one-on-one tutoring, and if both groups would demonstrate greater learning gains from pretest to posttest than students in business-as-usual classrooms. 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 the experimental condition with group discussions and Marni and 69 were received individual tutoring with Marni. When compared with the control group, effect sizes for were d *= 0.48 for the Group condition and* d *= -0.51 for the Individual condition. A two-way ANOVA suggested a main effect for tutoring group,* F *= 16.8, df (4,1171),* p *< .0001. In general, students reported benefiting from listening to one another, and the small group interactions, even though they sometimes did not agree with the group answer. We conclude our findings with a vision for a next generation of science tutoring using technology to help all students learn science using technology.*

Keywords: virtual, tutoring, intelligent, learning, student, multimedia

DO NOT CITE - Submitted for publication

Highlights

- The lack of discourse in U.S. classrooms is especially puzzling, given compelling evidence that effective programs have been developed in which teachers engage students in discourse leading to significant gains in student motivation and learning.
- Students in one condition engaged in a series of one-on-one tutoring sessions with a virtual science tutor, Marni. Following the multimedia presentations, Marni engaged students in question-answer interactive dialogs, in which students constructed explanations in response to Marni's questions. Students in a second condition interacted with Marni in small groups, of two to four students.
- 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 expert human tutors.
- Students in both one-on-one and small group 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. The magnitude of the learning gains was similar to those observed in previous investigations.
- Students reported that they enjoyed arguing with their peers, believed they achieved a deeper understanding of the science as a result of the discussions, even though most students reported that they sometimes disagreed with the answer presented to the virtual tutor following the discussion

DO NOT CITE - Submitted for Publication

1. Introduction

In this article, we describe a study in which we developed and evaluated a program to help students gain proficiency in scientific discourse. 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: "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."

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 facilitated by a virtual tutor. Students were presented with a set of narrated multimedia science presentations, and engaged in conversations about the science. This corresponded directly with the main themes of this special issue on children's speech: engaging early learners in spoken dialogs with both virtual and human teachers, with the ultimate goal of improving student motivation and achievement through productive discourse and argumentation.

The article is organized as follows. Section 1 provides an introduction to, overview of, and scientific rationale for the study. Section 2 describes the two experimental conditions used in 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 research, and briefly 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. Following the multimedia presentations, Marni engaged students in question-answer interactive dialogs, in which students constructed explanations in response to Marni's questions. Students in a second condition interacted with Marni in small groups, of two to four students. Students in small groups received the same multimedia presentations, but were asked to discuss each question before one of the students produced a spoken answer for the virtual tutor (VT), Marni. We solicited students' opinions, asked students in each condition about their learning experiences, and compared learning gains of students in each condition to students in control classrooms, who received the similar classroom instruction, but did not receive supplemental tutoring. Our main research questions were:

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

DO NOT SITE - Submitted for Publication

1.2. Scientific Foundations

1.1.1 Theory and Research on Scientific Discourse and Argumentation

Historically, research in argumentation and collaborative discourse acknowledges the strong influence 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 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 in speaking" (Roth, 2014), in online Abstract). We embrace this emphasis as a main focus of our research in helping children become comfortable using words as they express, support, defend, reflect on, and modify their ideas about science. The seminal writings Vygotsky and Bakhtin have had a profound influence on subsequent research in discourse and argumentation, including Wells' (1997; 2000; 2008) research on dialogic inquiry.

The past 25 years have witnessed remarkable growth in research on discourse and argumentation in education. Kuhn (1993, 2000) has argued that "a conception of science as argument has come to be widely advocated as a frame for science education." Support for 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; Harris, McNeill, Lizotte, Marx, & Krajcik, 2006; Kulatunga & Lewis, 2013; Kulatunga, Moog, & Lewis, 2013; McNeill, 2011; Nussbaum, Sinatra, & Poliquin, 2008; Sampson, Grooms, & Walker, 2009; Schworm & Renkl, 2007; Simon, Erduran, & Osborne, 2006; Voss & Means, 1991). Evidence from these studies indicates that argumentation can be improved by providing professional development to teachers or knowledgeable students (Berland & Reiser, 2009; Bricker & Bell, 2008, 2014; de Jong, Linn, & Zacharia, 2013) 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 (e.g., Kulatunga & Lewis, 2013; Kulatunga et al., 2013).

There is strong evidence that having students communicate with each other about science concepts taught during classroom instruction improves learning. 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, or *interactive classes* where teachers stopped their lectures to ask students to discuss questions. Students in interactive classes achieved 48% gains relative to 24% gains in classrooms where teachers lectured but did not ask questions.

Black & Wiliam's (2009) synthesis of over 250 research studies indicated that administering formative assessments to students, and providing and teachers and students with feedback on their performance, produced effect sizes of 0.40 to 0.70 on standardized tests, relative to students in classrooms who were not administered formative assessments. Students benefit from

DO NOT CITE Submitted for Publication

feedback on their understandings of the science they are learning, and teachers benefit from information that informs their instruction.

A meta-analysis by Chi (2009) indicated 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). In two studies testing the ICAP hypothesis, that Interactive instruction is more effective than Constructive, Active or Passive instruction (Menekse, Stump, Krause, & Chi, 2013) obtained strong support for ICAP hypothesis; interactive instruction led to higher scores on deep reasoning questions relative to constructive, active or passive instructional methods.

In sum, despite the paucity of discourse in most classrooms, there is compelling evidence that when teachers are trained to initiate and manage classroom conversations in which students share, compare, and modify their ideas, they become more engaged and excited about learning, and overall achievement improves.

1.1.2 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 (Martin Nystrand & Gamoran, 1991). Over a period of two years, trained observers paid four visits to each of 58 eighth grade and 54 ninth grade English classes in parochial and public schools in rural, suburban and urban settings. The observers recorded the amount of time spent in different instructional activities, and recorded and coded over 23,000 teacher and student questions for a set of variables contrasting monologic and dialogic dialogs. Questions were coded for *authenticity* (a question is authentic if the answer is not known in advance by the asker) and *uptake*, the incorporation of previous answers 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"(Martin Nystrand, 1997) (p. 44). On average, there was less than 50 seconds of discussion per class in eighth grade and 20 seconds per class in ninth grade. Approximately 60% of all classes had no discussions, the single classroom with the most discussion averaged 2 minutes. Interestingly, there was a significant positive correlation between the amount of discourse in individual classrooms and student achievement in language arts. These were results were replicated in a second large scale study (Nystrand, Gamoran, Kachur, & Prendergast, 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" (pg. XX).

The lack of discourse in U.S. classrooms is especially puzzling, given compelling evidence that effective programs have been developed in which teachers engage students in discourse leading to significant gains in student motivation and learning (P. Karen Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009; Soter et al., 2008). One of these programs, Questioning the Author (Beck, Isabel, McKeown, & Margaret, 2006), identified QtA as one of two approaches out of the nine examined that is likely to promote high-level thinking and comprehension of text.

Relative to control conditions, QtA showed effect sizes of 0.63 on text comprehension measures, and of 2.5 on researcher-developed measures of critical thinking/reasoning (P. K. Murphy & 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

DO NOT CITE. Submitted for Publication

indicators of productive discussions likely to promote learning and comprehension of text (Gamoran & Nystrand, 1991; P. Karen Murphy et al., 2009). QtA provides the basis of the dialog moves used by the virtual tutor Marni, and by classroom teachers in the studies described below.

1.3.1 Benefits of Individualized Instruction: Human Tutoring

Over three decades of research indicate 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 classroom instruction. Evidence that tutoring works has been obtained from dozens of well-designed research studies and meta- analyses (Cohen, Kulik, & Kulik, 1982) and positive outcomes were obtained in large-scale tutoring programs (Slavin & Madden, 1989; Topping & Whiteley, 1990).

Factors that contribute to the effectiveness of individualized instruction, measured by gains in student achievement attributed to tutoring, include aligning tutoring content to classroom instruction, asking students authentic questions, scaffolding learning following questions with hints, follow-on questions and media to stimulate reasoning and help students' build on prior knowledge, and having students continually generate explanations in response to deep reasoning questions (Butcher, 2006; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Craig, Gholson, Ventura, & Graesser, 2000; Driscoll et al., 2003; King, 1989; King, Staffieri, & Adelgais, 1998; Palinscar & Brown, 1984; Pine & 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" (pg. 418).

1.3.2 Benefits of Individualized Instruciton: Intelligent Tutoring Systems

Research in intelligent tutoring systems address a 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 similar to that provided by an expert human tutor. Advances in Intelligent Tutoring System (ITS's) during the past 15 years have resulted in systems that produce learning gains comparable to those achieved with human tutors. A recent review by Van Lehn (2011) compared learning gains in studies in students received human tutoring or interacted with an ITS, in tasks that required students to engage in problem solving and constructing 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 ITS's was d=0.76. Van Lehn (2011) concluded that "intelligent tutoring systems are nearly as effective as human tutoring systems" (pg. 197).

1.3.3 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 expert human tutors (Ward et al., 2011, 2013).. To our knowledge, MyST is the only system to date that supports spoken tutorial dialogs with children (aged 7 to 10). MyST dialogs are aligned to to science concepts encountered in small-group classroom science investigations, using the Full Option Science System (FOSS) program. FOSS is a kit-based, inquiry-based program, used by over 1 million students in 100,000 classrooms in the U.S. (FOSS, 2007). During evaluations of MyST, students engaged in conversations with the virtual tutor Marni soon after they participated in a classroom science investigation. The tutorial dialog session was designed to

DO NOT CITE. Submitted for publication

help students construct explanations about the science covered in the classroom science investigation.

A defining feature of the tutorial dialogs is that Marni asked students open-ended questions about science presented in media—static illustrations, silent animations, or interactive simulations; questions, like "What's going on here?" or "What would happen if…?" Sessions lasted between 15 and 20 minutes, with Marni asking questions and presenting media with the goal of stimulating students to reason about the science and construct accurate explanations. Summative evaluations of My Science Tutor demonstrated learning gains of about half a grade, comparable to human tutoring, relative to students who did not receive tutoring by a virtual or human tutor as a supplement to classroom instruction (Ward et al., 2011, 2013).

What we describe below differed from previous MyST evaluations in that students were presented with narrated multimedia science presentations, and were then asked questions about the science in the presentations. Hereafter, we refer to spoken dialogs in our previous MyST studies (Ward et al. 2011; 2013) as MyST-SLS to distinguish the dialog approach used in MyST-SLS (silent animations) from the tutoring strategy used in the current study, which incorporates both narrated multimedia presentations and dialogs (MyST-MP&D).

2. Investigating One-on-One and Small Group Conversations with a Virtual Tutor

In MyST-MP&D students were presented with a narrated animation that posed a science question or problem, followed by Marni asking questions about the science question or problem. Students were then presented with a second narrated multimedia presentation, which provided an explanation of the science. This presentation was followed by a series of questions designed to elicit explanations. Students were randomly assigned to one-on-one or small group tutoring conditions. Students in each condition were presented with the same multimedia presentations, and the same questions following the presentations. At the end of end of sessions, students engaged in a short (2 to 5-minute) dialog with Marni. During this dialog, Marni asked students the same deep reasoning question, but follow-on questions depended on students' responses. This final dialog was a greatly reduced version of MyST dialogs used in our prior studies, described in Ward et al. (2011; 2013), aligned to science investigations in FOSS Measurement, and Magnetism and Electricity. For ease of exposition, we use examples from the Measurement sessions to describe the study; results are reported for both modules.

2.1.1. *Sequence of MyST-MP&D Activities*

2.1.1.1. Title Screen

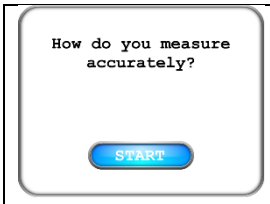
Each MyST-MP&D session began with a title screen that presented a deep reasoning question. In all cases, the printed question was read aloud by the virtual tutor. Examples included: What do magnets stick to? What is an electrical circuit? How can we measure length (volume, mass, temperature) and get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

2.1.1.2. Engaging Real-life Scenario

DO NOT CITE, submitted for publication

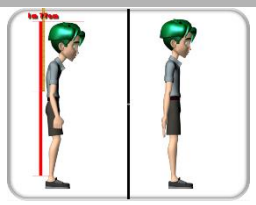
The first multimedia presentation was a narrated animation that introduced the science. It associated the science with materials and situations likely to be familiar to most or all of the students, based on classroom instruction. The Scenario was followed by an open-ended question by Marni, like "What's 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 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 and the deep reasoning question introduced on the title screen of the MYST-MP&D session (see Figures 1 and 2).

Figures



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 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"

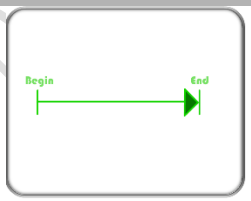
"What's going on here?"

"How do you measure accurately?

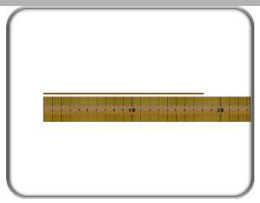
Students can decide to replay the narrated multimedia presentation, or press continue, at which Marni asks an open-ended question, such as "Why do you think Jill got differences answers each time she measured Jack?" After students have answered the question, which is recorded, Marni produces a prompt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

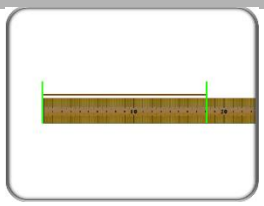
Figure 1. Sample media.

EXPLANATION



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

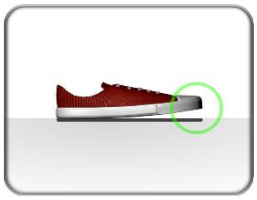




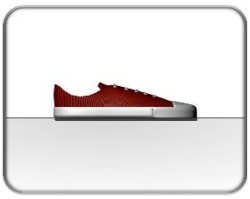
"It is also very important to make sure things are flat and lined up with your meter stick."

Do Not Cite - Submitted for Publication

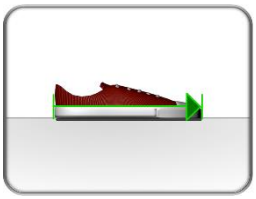
"Take this shoe for example."



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

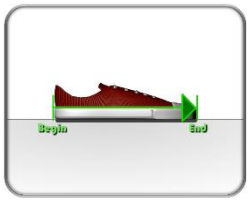


"In order to get a good measurement, we first have to make sure that the object we are measuring is as flat as possible."

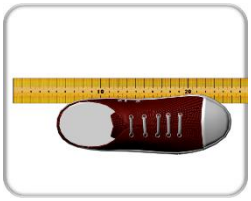


"Then we start our measurements at the back of the shoe

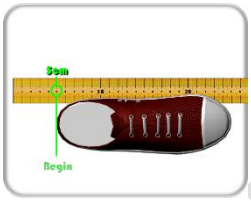
And measure all the way to the tip of the shoe."



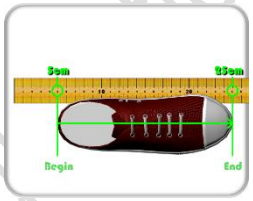
"These two spots are the beginning and end of our shoe"



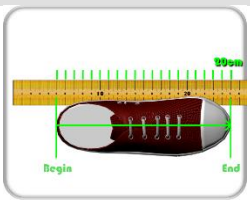
"But we need our meter stick to make an actual measurement."



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

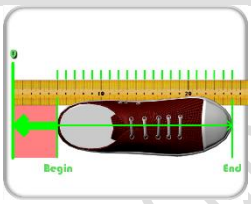


"…and the tip of our shoe is at the 25 centimeter mark. Does that mean our shoe is 25 cm long?"

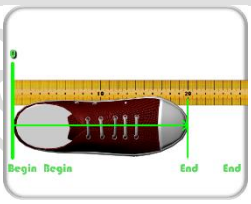


" Well, when we actually count the centimeters starting at the back of the shoe and ending at the tip of our shoe….we count twenty centimeters, not twenty-five."

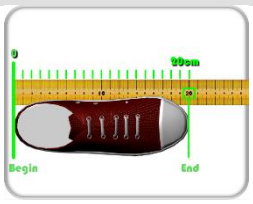
"Oh, instead of counting the units in-between, is there an easier way?"



"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."

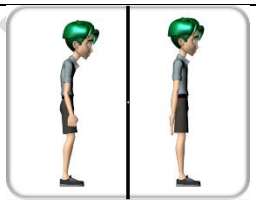


"And see what happens? Our shoe starts at the zero mark and ends at the twenty mark."

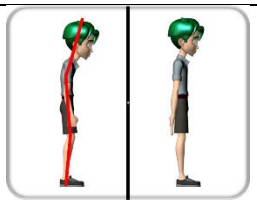


"We measured the shoe as being twenty centimeters long. That's the same measurement that we got before. Perfect!"

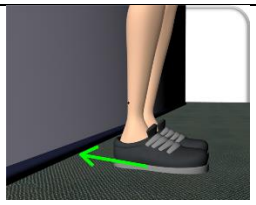
Figure 2. Explanation in sample media.



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



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



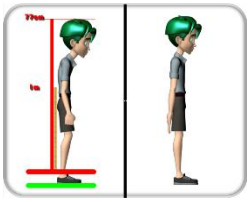
"And my first were not close to the wall, so I was not flat against the wall."

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

NOT CITE Submitted for Publication

measurements."

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

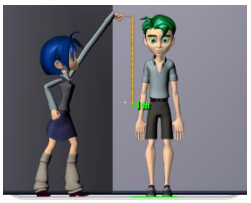


"And remember the first time the meter stick was not down on floor by your feet. It was up by your calf above your ankle. See, I started at the wrong place."

"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 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."

Figure 3. Explanation in sample media.

DO NOT CITE - Submitted for Publication

2.1.1.3. Multimedia Science Explanation

Students were presented with a multimedia presentation that explained the science. The design of the narrated multimedia presentations was informed by 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 multimodal mental models that integrate 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 presented, so students' had time to process each of key points in the presentations. Figures 2 and 3 present an example of a multimedia explanation for measurement.

2.1.1.4. Formative Assessment – Multiple Choice Questions (MCQs)

After the multimedia presentation was 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.

2.1.1.5. 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.

2.1.1.6. Spoken Dialogs with Marni

Each session concluded with a spoken dialog with Marni, lasting less than 5 minutes. These were truncated versions of 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.

2.1.2. Results

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 gain more from pretests to posttest than students receiving no tutoring.*

We did not expect statistically significant differences learning gains between students 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 previously.

DO NOT CITE. Submitted for Publication

2.1.2.1. Research Design and Data Collection 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-MP&D 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' con versations, as discussed below (Table 1).

Table 1. Means, Standard Deviation, Pre/Post Average and Scale for FOSS-ASK tests.

Module		Pre (raw)	Post (raw)	Pre/Post Average	Scale
Measurement	Mean	27.8	44.6	36.2	9 – 63
	SD	8.7	7.9	8.3	
M & E	Mean	21.2	29.6	25.4	7 – 39
	SD	7.4	8.8	8.1	

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-MP&D by answering questions verbally, or by choosing multiple choice answers.

2.1.2.2. 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

DO NOT CITE - Submitted for publication

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 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.

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 2.

Table 2. Mean, SD, N and Effect Size for tutoring groups 2011 and 2012.

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

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 0.89 to 0.94, with averages for pre and post 0.91 and 0.94. Internal reliabilities (Cronbach's Alpha) were lower, ranging from $a = 0.66$ to $a = 0.87$ for both pre and post versions of the assessments, with averages for pre = 0.78 and post = 0.78. Internal reliability varied for each module. Scores used for outcome analysis were the averages across both raters.

Table 3. Mean, SD, N and Effect Size for tutoring groups 2011 and 2012.

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

DO NOT CITE - Submitted for Publication

2.1.2.3. Results

When compared with the control group, effect sizes for were ***d* = 0.48** for the Group condition and ***d* = -0.51** for the Individual condition. These were both close to what we found for the year before for individual tutoring with MyST-SLS (*d* = 0.51) and for human tutors (*d* = 0.65). 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 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 4).

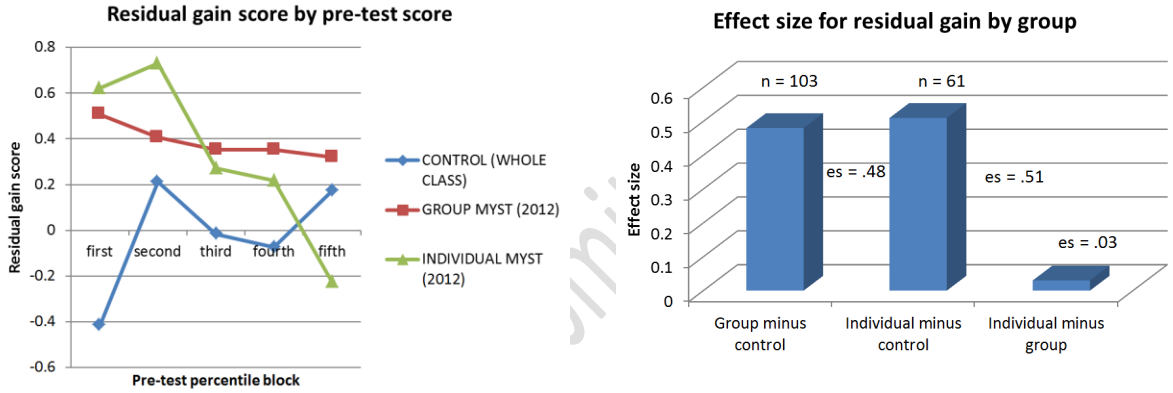


Figure 4. Residual gain scores over time and experimental groups.

Comparisons with previous treatment groups from 2010-2011. A two-way 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-MP&D. 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 portable digital assistant (PDA) and electronic data was imported into an Excel database.

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.

2.1.2.3. 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

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

DO NOT CITE - Submitted for Publication

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

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

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

Links between FOSS-ASK assessments and types of responses in small groups. We also wanted to know if gain on the FOSS-ASK assessment was related to the frequency and duration of discussions. The average amount of time spent by students in interactive discussions was correlated (r = 0.23) 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.

2.1.2.4. Students 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-SLS 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. These are shown in Figures 5-6.

DO NOT CITE. Submitted for Publication

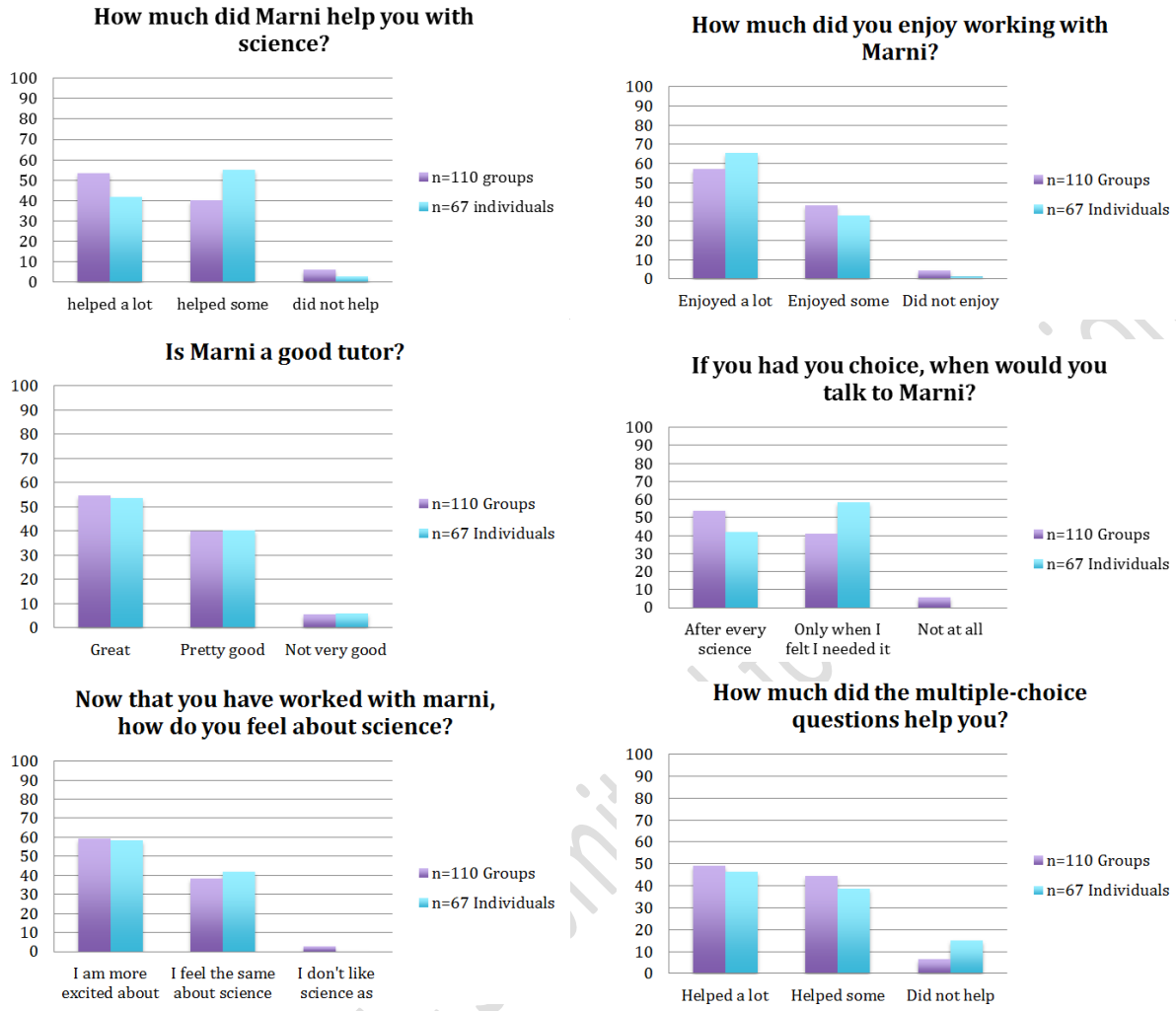


Figure 5. Students' response to surveys in One-On-One condition.

DO NOT CITE

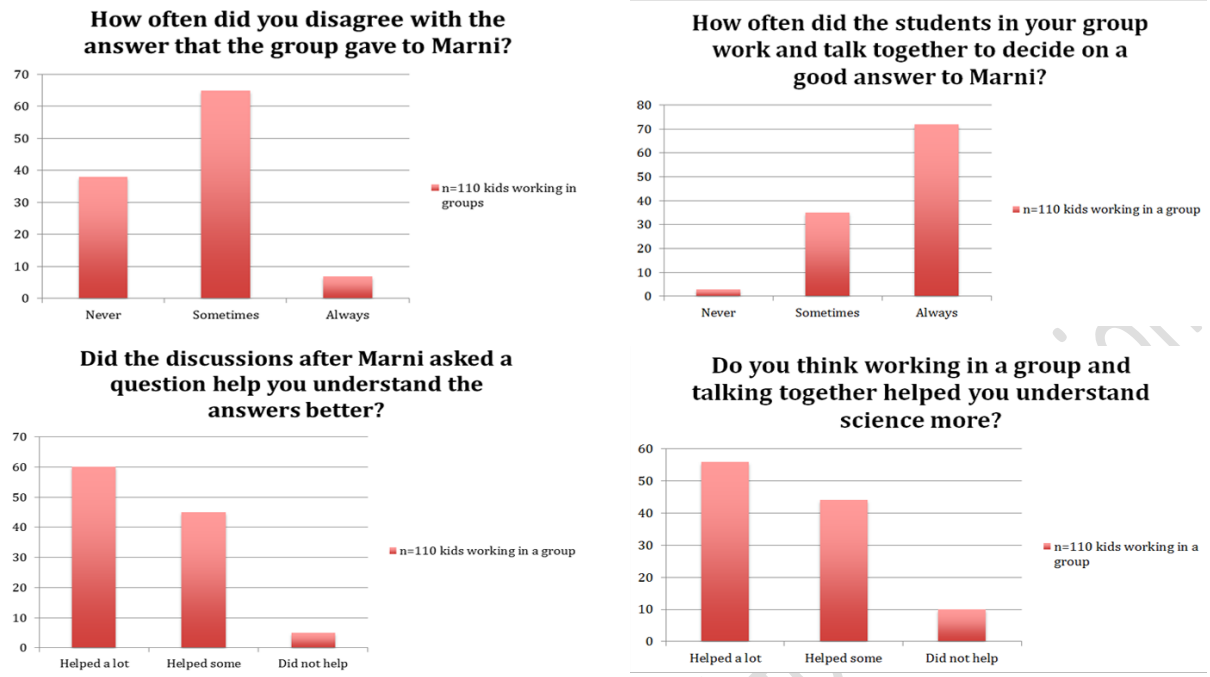


Figure 6. Students' responses to surveys in Small Groups condition.

3. Discussion

This study investigated student experiences and learning gains following one-on-tutoring sessions with a virtual science tutor and small group tutoring sessions in which students were encouraged to discuss answers to questions posed by the virtual tutor before one of the students presented a group answer to the tutor. The results indicated that learning gains were comparable for the two groups, and students in both groups had generally positive experiences.

The magnitude of the learning gains was similar to those observed in previous investigations (see Ward et al., 2011; 2013). This investigation addressed issues with how individual and groups of students interacted with multimedia, spoken dialogues with a virtual tutor about science presented—which was the special topic of the corresponding Journal Issue. The main difference between the one-on-one tutoring sessions in the prior studies, and in the current study, was that students in our prior studies engaged in spoken dialogs during the entire tutoring session (rather than just part of it), as they explained science presented to them, answered questions about the science content, and conducted follow-up activities and questions posed by the virtual tutor. The assumption in these dialogs was that students had acquired sufficient prior knowledge during classroom science instruction to build on this knowledge with the virtual tutoring sessions. This would in turn enable them to engage in meaningful discussions about science content with a virtual tutor that scaffolded learning through questions and multimedia presentations.

In the current study, students in both one-on-one and small group 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 prior packages with virtual tutors, these were more "self-contained," in that they presented sufficient

DO NOT CITE. Submitted to

science content to facilitate subsequent discussion and construction of explanations. Overall, students that received the tutoring demonstrated greater learning gains than nontreatment students. Further research is required to understand the relative contributions of incorporating narrated multimedia problem scenarios and science explanations into dialogs. At this point, we concluded that both approaches produced comparable learning gains, when compared to students in control classrooms who received similar classroom instruction, but did not receive supplementary tutoring. We also concluded that the relationship between student achievement and particular aspects of the virtual tutoring any student received was complex and required additional investigations.

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 our 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 a tutor's questioning (scaffolding). Students reported that they enjoyed arguing with their peers, believed they achieved a deeper understanding of the science as a result of the discussions, even though most students reported that they sometimes disagreed with the answer presented to the virtual tutor following the discussion. Future work will aim to extend the capabilities of the virtual science tutor, to foster effective discourse and argumentation, and to better understand the complex relationship between the types of multimedia instructional support and changes in student knowledge.

4. Conclusion

This study demonstrated that students who received one-on-one tutoring with a virtual tutor, and students who engaged in small group discussions following questions posed by the virtual tutor, can achieve the same significant learning gains, relative to students who do 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 to learn argue effectively, defend their arguments, and modify their ideas as they co-construct accurate science explanations with their peers.

We envision a future wherein students sit around a table, each with their own tablet or laptop computer. Students have a clear view of each other and the virtual tutor, whose face appears on her own computer screen at the head of the table. Each student wears lightweight headphones with an attached, noise-cancelling microphone. Each student views and interacts with media on their computer, and uses their laptop to answer multiple-choice questions or interact with simulations. Students talk to the virtual tutor by pushing and holding the spacebar on their computer, or button on their tablet which causes a light to go on, indicating which student has the floor.

During small group discussions with the tutor, speech recognition and natural language processing systems construct a semantic representation of each student's expressed understandings, gaps in their knowledge, and misconceptions. The tutoring system also builds a representation of the distributed knowledge of the group; i.e., which semantic frames have been filled during the conversation by the different members of the group. By creating a unified

DO NOT CITE - Submitted for Publication

representation of the group's knowledge from representations of each individual student's expressed knowledge, the system is able to identify conflicting beliefs and commonly missing elements of the science content. It will identify parts of students' arguments that are based on their correct knowledge and misconceptions of science. It will ask follow-up questions to individual students or the group that target students' expressed misconceptions, and that focus on differences in the way students think about similar scientific phenomena. In doing this, it will foster discourse and argumentation in which students listen to each other, consider each other's arguments, modify their ideas, and work together to construct explanations that result in a deeper understanding of the content for the entire group.

Furthermore, by building representations of each student's understanding, while simultaneously tracking which components of a complete explanation the group has achieved, the tutor can direct questions to specific students to help them resolve conflicting explanations through effective argumentation when it behooves both that individual and the group based on the importance of the content, or the utility of the animations or questions associated with discussing that content with the individual student in the presence of the entire group.

We envision the tutor using dialog moves based on principles of QtA, to model scientific discourse, and address questions to individual students to foster effective argumentation. A core principle of the system and its responses to students' interactions 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 (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 tutor models the process of listening carefully, extracting meaning, and making connections between students' ideas using established dialog moves which paraphrase or elaborate on student's ideas. The tutor models scientific discourse by rephrasing what a student said. When Roger says, "It flows in one direction from negative to positive." The tutor may respond, "I think you said electricity flows from the positive pole to the negative pole of the battery". The tutor then asks a question to the same student, or to another student *(e.g., Maria, do you agree with what Roger said?)* The tutor can also pose questions to the entire group (e.g., *So what's going on here?* or, *Why don't you discuss the different explanations Maria and Roger presented?*).

We hypothesize, based on the results of this study, that our current technologies are fully capable of facilitating such group discussions. Current speech recognition, language processing, and dialog modeling technologies demonstrably support graceful and effective dialogs between virtual tutors and students (Ward et al., 2011; 2013). Furthermore, we believe that is relatively straightforward to extend the representations used in MyST-SLS for one-on-one tutoring to model the expressed understandings of each member of a group, while maintaining a model of the groups' combined knowledge. Moreover, the dialog modeling currently applied to individual students' spoken answers in MyST-SLS, can be extended to support conversations with multiple students, thereby faithfully generating QtA dialog moves for groups of students, which have been demonstrated effective in fostering discourse in 100's of classrooms (Beck, Isabel, McKeown, & Margaret, 2006).

Lastly, research in the tutoring literature has suggested that all students who participate in small group conversations may benefit from listening to other students as they argue about science, without necessarily contributing to the discussion. Recent studies indicated that students, who observed tutorial dialogs in which deep reasoning questions were asked between tutoring turns of

DO NOT CITE - Submitted for Publication

students they were watching, achieved learning gains equivalent to students who actively participated in the tutoring sessions (Chi, Roy, & Hausmann, 2008; Craig, Gholson, Ventura, & Graesser, 2000). These studies suggest that a new generation of intelligent tutoring systems that can listen to and foster discourse and argumentation among groups of students, may benefit all participants.

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

DO NOT CITE Submitted for Publication

References Cited

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

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

Beck, Isabel, McKeown, & Margaret. (2006). *Improving comprehension with Questioning the Author: A fresh and expanded view of a powerful approach*. New York, NY: Scholastic.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26–55.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 5–31.

Bloom. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. *Educational Researcher*, *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. *Science Education*, *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. *Journal of Research in Science Teaching*, *51*(3), 260–285.

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology*, *98*(1), 182–197.

Chi, M. T. H. (2009). Active-constructive-interactive: a conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *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. *Cognitive Science*, *13*(2), 145–182.

Chi, M. T. H., & Menekse, M. (2015). Dialogue patterns that promote learning. In L. B. Resnick, C. Asterhan, & S. N. Clarke (Ed.), *Socializing intelligence through academic talk and dialogue* (pp. 263–273).

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: insights about human tutoring effectiveness from vicarious learning. *Cognitive Science*, *32*(2), 301–341.

Chin, C., & Osborne, J. (2010). Supporting Argumentation through Students' Questions: Case Studies in Science Classrooms. *Journal of the Learning Sciences*, *19*(2), 230–284.

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*(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. *International Journal of Artificial Intelligence in Education*, *11*, 242–225.

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

Driscoll, D., Craig, D., S., Gholson, B., … A. (2003). Vicarious learning: Effects of overhearing dialog and monolog-like discourse in a virtual tutoring session. *Journal of Educational Computing Research*, *29*, 431–450.

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

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

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*, 487–493.

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

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*, 67–88.

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.

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

King, A., Staffieri, A., & Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology*, *90*(1), 134.

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

Kuhn, D. (2000). Metacognitive Development. *Current Directions in Psychological Science*, *9*(5), 178–181.

DO NOT CITE - Submitted for Publication

Kulatunga, U., & Lewis, J. E. (2013). Exploration of peer leader verbal behaviors as they intervene with small groups in college general chemistry. *Chemical Education Research and Practice*, *14*(4), 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.

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

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.

Menekse, M., Stump, G. S., Krause, S., & Chi, M. T. H. (2013). Differentiated Overt Learning Activities for Effective Instruction in Engineering Classrooms. *Journal of Engineering Education*, *102*(3), 346–374.

Murphy, P. K., & Edwards, M. N. (2005). What the studies tell us: A meta-analysis of discussion approaches. In M. Nystrand (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. *Journal of Educational Psychology*, *101*(3), 740–764.

National Research Council. (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8* (Committee on Science Learning Kindergarten through Eighth Grade). Washington D.C.: The National Academies Press. Retrieved from http://www.nap.edu/catalog/11625.html

Nussbaum, E. M., Sinatra, G. M., & Poliquin, A. (2008). Role of Epistemic Beliefs and Scientific Argumentation in Science Learning. *International Journal of Science Education*, *30*(15), 1977–1999.

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. (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. *Research in the Teaching of English*, *25*(3), 261–290.

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. *Cognition and Instruction*, *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. *Cognition and Instruction*, *18*(1), 35–51.

DO NOT CITE - Submitted for Publication

Roth, W.-M. (2013). An Integrated Theory of Thinking and Speaking that Draws on Vygotsky and Bakhtin/Vološinov. *Dialogic Pedagogy: An International Online Journal*, *1*(0), 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*(8), 1049–1083.

Sampson, V., Grooms, J., & Walker, J. (2009). Argument-driven inquiry. *The Science Teacher*, *76*(8), 42–47.

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

Simon, S., Erduran, S., & Osborne, J. (2006). Learning to Teach Argumentation: Research and development in the science classroom. *International Journal of Science Education*, *28*(2-3), 235–260.

Slavin, R. E., & Madden, N. A. (1989). What Works for Students at Risk: A Research Synthesis. *Educational Leadership: Journal of the Department of Supervision and Curriculum Development, 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. *International Journal of Educational Research*, *47*(6), 372–391.

Sullins, J., Craig, S. D., & Graesser, A. C. (2010). The influence of modality on deep-reasoning questions. *International Journal of Learning Technology*, *5*(4), 378.

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

VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, *46*(4), 197–221.

Voss, J. F., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction*, *1*(4), 337–350.

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

Vygotsky. (1987). *Thinking and speech*. New York, NY: Plenum.

Ward, W., Cole, R., Bolaños, 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 Transactions on Speech and Language Processing (TSLP)*, *7*(4), 18.

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*(4), 1115–1125.

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.

DO NOT CITE - Submitted for Publication

Wells, G. (1997). A sociocultural perspective on classroom discourse: Appendix A. Coding scheme for the analysis of classroom discourse. *The Encyclopedia of Language and Education*, *3*, 3–5.

Wells, G. (2008). Learning to Use Scientific Concepts. *Cultural Studies of Science Education*, *3*(2), 329–350.

DO NOT CITE - Submitted for Publication