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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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.

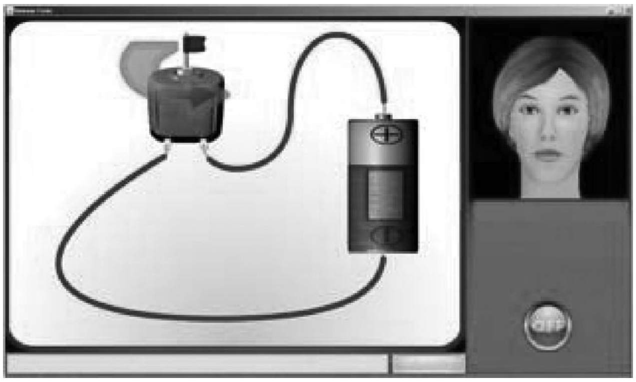

Figure 1. My Science Tutor (MyST) screen layout.

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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

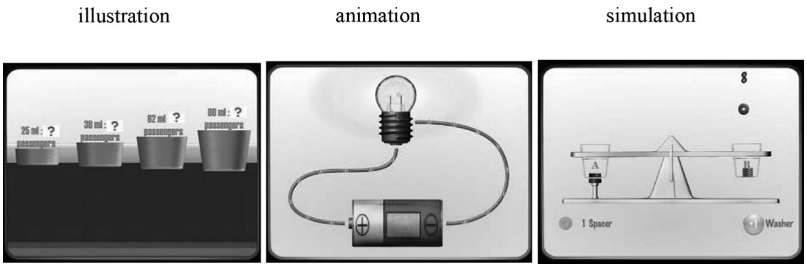

Figure 2. Media types.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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 classrooms, 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 stu-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

dents 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 $\alpha = .60$ to $\alpha = .89$ 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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.