Recent Advances in Conversational Intelligent Tutoring Systems

Vasile Rus, Sidney D'Mello, Xiangen Hu, Arthur C. Graesser

■ *We report recent advances in intelligent tutoring systems with conversational dialogue. We highlight progress in terms of macro- and microadaptivity. Macroadaptivity refers to a system's capability to select appropriate instructional tasks for the learner to work on. Microadaptivity refers to a system's capability to adapt its scaffolding while the learner is working on a particular task. The advances in macro- and microadaptivity that are presented here were made possible by the use of learning progressions, deeper dialogue, and natural language-processing techniques, and by the use of affect-enabled components. Learning progressions and deeper dialogue and natural language-processing techniques are key features of DeepTutor, the first intelligent tutoring system based on learning progressions. These improvements extend the bandwidth of possibilities for tailoring instruction to each individual student, which is needed for maximizing engagement and ultimately for learning.*

The vision of one-on-one human tutoring being the most effective solution to instruction and learning has attracted the attention of many for decades. Encouraged by the effectiveness of one-on-one human tutoring (Bloom 1984), computer tutors that mimic human tutors have been successfully built with the hope that a computer tutor could be available to every child with access to a computer. An extensive review of tutoring research by VanLehn (2011) showed that computer tutors are as effective as human tutors. VanLehn reviewed studies published between 1975 and 2010 that compared the effectiveness of human tutoring, computer-based tutoring, and no tutoring. The conclusion was that the effectiveness of human tutoring is not as high as it was originally believed (effect size $d = 2.0$) but much lower ($d = 0.79$). The effectiveness of computer tutors ($d = 0.78$) was found to be as high as the effectiveness of human tutors. So, there is something about the one-on-one connection that is critical, whether the student communicates with humans or computers. Graesser, Person, and Magliano (1995) argued that the remedial part of tutorial interaction in which tutor and tutee collaboratively improve an initial answer to a problem is the primary advantage of tutoring over classroom instruction. Chi, Siler, and Jeong (2004) advanced a related hypothesis: tutoring enhances students' capacity to reflect iteratively and actively on domain knowledge. Furthermore, one-on-one instruction has the advantage of engaging most students' attention and interest as opposed to other forms of instruction such as lecturing or monologue in which the student may or may not choose to pay attention (VanLehn et al. 2007).

Intelligent tutoring systems (ITSs) with conversational dialogue form a special category of educational technologies. These conversational ITSs are based on explanation-based constructivist theories of learning and the collaborative constructive activities that occur during human tutoring. They have proven to promote student learning gains up to an impressive effect of 1.64 sigma when compared to students learning the same content in a canned text remediation condition that focuses on the desired content (VanLehn et al. 2007). Of course, not all meta-analyses and research reports are so rosy, as indicated in the Dynarsky et al. (2007) report on commercial learning systems in K–12 environments and also some experimental conditions investigated by VanLehn and colleagues (2007). For instance, when intermediate students are given intermediate content and when novice students are given novice content, then tutoring is no better than canned text remediation. The true impact of conversational tutoring on learning is still not settled empirically.

The conventional wisdom of the last decade has speculated that as interactivity of tutoring increases, the effectiveness of tutoring should keep increasing. However, VanLehn (2011) reported that as interactivity of tutoring increases, the effectiveness of human and computer tutors plateaus. VanLehn's finding challenged ITS developers to find new approaches to further increase computer tutors' effectiveness. Indeed, novel approaches to ITS development, such as the ones presented here, are needed to push their effectiveness beyond the interactivity plateau.

There are several aspects of state-of-the-art conversational ITSs that may explain the plateau in their effectiveness. First, they do not emphasize macroadaptation through selection of learner-tailored content and tasks, which is needed when students begin tutoring sessions with different backgrounds. Second, conversational ITSs rely on dialogue and language-processing algorithms to guide the interaction between the tutor and tutee. The quality of these algorithms has a direct impact on core ITS tasks such as summative and diagnostic assessment, that is, the detection and tracking of students' knowledge states, and providing formative feedback. Assessment and feedback are critical components of any tutoring system that is fully adaptive. There is room for improvement when it comes to dialogue and language processing in conversational ITSs. Third, existing conversational ITSs emphasize mostly cognitive aspects when providing microadaptation. Other aspects of learning, such as affect and motivation, are typically not considered to guide microadaptation.

In this article, we present advances in conversational ITSs that address the above weaknesses and that have been proposed over the last decade, since the 2001 *AI Magazine* report by Graesser et al. (2001) on the state of the art in conversational ITSs. We highlight (a) the addition of the learning progressions (LPs) framework to increase macroadaptivity (Corcoran, Mosher, and Rogat 2009); (b) deeper dialogue and natural language processing to increase accuracy of student input assessment and quality of tutor feedback (Rus and Graesser 2006; Rus and Lintean 2012); this improves macro- and microadaptivity of ITSs; and (c) modeling students' affective states in addition to their cognitive states (D'Mello et al. 2013); affect-enabled components improve the microadaptivity capabilities of ITSs. It should be noted that LPs can have a significant impact on microadaptivity as well but we do not emphasize these aspects in this article. Furthermore, affect-enabled components may affect macroadaptivity as well but we do not emphasize these aspects either.

Taken together, these advances presented in this article form one promising approach to increasing the effectiveness of tutoring systems beyond the interaction plateau (VanLehn et al. 2007, VanLehn 2011).

Intelligent Tutoring Systems with Animated Conversational Agents

Conversational ITSs have several advantages over other types of ITSs. They encourage deep learning as students are required to explain their reasoning and reflect on their basic approach to solving a problem. Conceptual reasoning is more challenging and beneficial than mechanical application of mathematical formulas. Furthermore, conversational ITSs have the potential of giving students the opportunity to learn the language of scientists, an important goal in science literacy. A student associated with a more shallow understanding of a science topic uses more informal language as opposed to more scientific accounts (Mohan, Chen, and Anderson 2009).

The impact of conversational ITSs allegedly can be augmented by the use of animated conversational agents that have become more popular in contemporary advanced learning environments (Graesser et al. 2008). The animated agents interact with students and help them learn by either modeling good pedagogy or by holding a conversation with the learners. Both single agents and ensembles of agents can be carefully choreographed to mimic virtually any activity or social situation: curiosity, inquiry learning, negotiation, interrogation, arguments, empathetic support, helping, and so on. Agents not only enact these strategies, individually or in groups, but can also think aloud while they do so.

Examples of successful conversational ITSs are AutoTutor (Graesser et al. 2008), Why2 (VanLehn et al. 2007), CIRCSIM-Tutor (Evens and Michael 2005), and GuruTutor (Olney et al. 2012). DeepTutor,[1] described here, is an emerging conversational ITS.

We will focus next on AutoTutor, a conversational ITS described in the *AI Magazine* report by Graesser and colleagues (2001) and on VanLehn's (2006) two-loop framework for describing ITSs. The advances in conversational ITSs that we emphasize in this article can be best understood with respect to AutoTutor's basic dialogue and pedagogical framework and VanLehn's (2006) two-loop framework for describing ITSs.

AutoTutor Overview

AutoTutor is a conversational ITS that helps students learn science topics by holding a mixed-initiative dialogue with students in natural language (Graesser et al. 2008). The structure of the dialogue in both AutoTutor and human tutoring follows an expectation and misconception tailored (EMT) dialogue. EMT dialogue is the primary pedagogical method of scaffolding good student answers.

The dialogue moves and the problems AutoTutor can tutor on are stored in a curriculum script. The curriculum script is a knowledge structure employed by novice tutors that largely determines the content and flow of a tutoring session. AutoTutor promotes active construction of knowledge by providing explanations only when the learner is floundering. Moves at the beginning of the tutorial interaction for each new problem (that is, pumps and hints) provide less information to the student than later moves (that is, prompts and assertions).

The behavior of AutoTutor and that of any ITS, conversational or not, can be described using VanLehn's (2006) two-loop framework. According to VanLehn, ITSs can be described in broad terms as running two loops: the outer loop, which selects the next task to work on, and the inner loop, which manages the student-system interaction while the student works on a particular task. AutoTutor's outer loop is usually just an iteration over the set of existing tasks (which represent the main questions or problems; Graesser et al. [2003]). That is, all students go through the same set of tasks in the same order. Some level of adaption does exist as the set of tasks can be chosen based on known student conceptions and misconceptions. However, once the tasks are designed all students see the same tasks regardless of their individual differences. It should be noted that sometimes this one-size-fits-all approach of selecting training tasks is required by experimental design constraints, which dictate that all students must be exposed to the same content. Ideally, students should work on tasks that best suit their background. The use of LPs in the emerging conversational ITS DeepTutor will enable achieving this goal.

The inner loop of an ITS monitors students' performance through embedded assessment, updates its model of students' levels of understanding (that is, the student model), and uses the updated student model to provide appropriate scaffolding in the form of feedback and other scaffolds. In dialogue-based ITSs, embedded assessment relies heavily on language understanding algorithms as students' responses are natural language utterances. AutoTutor's language-processing algorithms rely on co-occurrence methods such as Latent Semantic Analysis (LSA; Landauer et al. [2007]), inverse weighted word frequency overlap, and regular expressions. As previously mentioned, another important aspect of the inner loop is to provide appropriate scaffolding while the student is working on a task, for example, correct a misconception immediately through appropriate feedback. We discuss in this article advances about how to better assess students' natural language inputs and how to better tailor scaffolding using affect-sensitive components while the student is working on a task.

The three major advances in conversational ITSs described in this article will have direct impact on both the outer and inner loops of ITSs and will lead to improvements in core tasks handled by ITSs: modeling the task domain, tracking students' knowledge states, selecting appropriate learning trajectories, and the feedback mechanisms. Advances in these core tutoring tasks will move state-of-the-art ITSs closer to implementing fully adaptive tutoring, which implies tailoring instruction to each individual student at both macro- and microlevel.

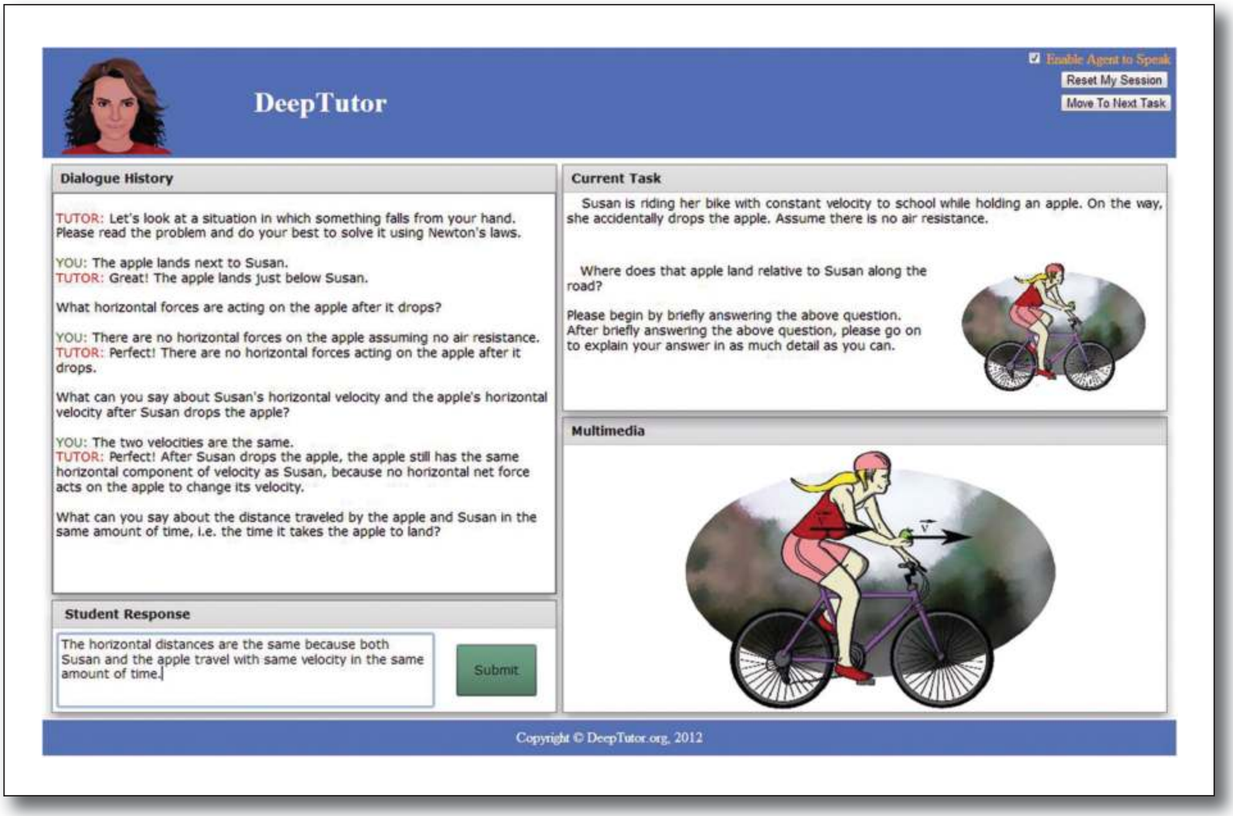
Advanced Macroadaptivity in DeepTutor

DeepTutor is a conversational ITS that is intended to increase the effectiveness of conversational ITSs beyond the interactivity plateau by promoting deep learning of complex science topics through a combination of advanced domain modeling methods, deep language and discourse processing algorithms, and advanced tutorial strategies. DeepTutor has been developed as a web service and a first prototype is fully accessible through a browser from any Internet-connected device, including regular desktop computers and mobile devices such as tablets. A snapshot of the learner view is shown in figure 1. DeepTutor currently targets the domain of conceptual Newtonian Physics but it is designed with scalability in mind (cross-topic, cross-domain). The spin-off project of AuthorTutor[2] aims at efficiently porting DeepTutor-like ITSs to new domains by investigating well-defined principles and processes as well as developing software tools that would enable experts to efficiently author conversational computer tutors across STEM disciplines. Another authoring tool, called SEMILAR (derived from semantic similarity toolkit; Rus et al. [2013]), is being developed as well to assist with authoring algorithms for deep natural language processing of student input in conversational ITSs.[3]

It is beyond the scope of this article to describe all the novel aspects of DeepTutor or related projects. Instead, we focus on two components of DeepTutor: domain modeling based on LPs and deeper dialogue and natural language processing.

Learning Progressions

LPs have been developed by the science education research community as a way forward in science education. The National Research Council (2001) report called for better descriptions of how students learn based on models of cognition and learning. Based on



Dialogue History

TUTOR: Let's look at a situation in which something falls from your hand. Please read the problem and do your best to solve it using Newton's laws.

YOU: The apple lands next to Susan.
TUTOR: Great! The apple lands just below Susan.

What horizontal forces are acting on the apple after it drops?

YOU: There are no horizontal forces on the apple assuming no air resistance.
TUTOR: Perfect! There are no horizontal forces acting on the apple after it drops.

What can you say about Susan's horizontal velocity and the apple's horizontal velocity after Susan drops the apple?

YOU: The two velocities are the same.
TUTOR: Perfect! After Susan drops the apple, the apple still has the same horizontal component of velocity as Susan, because no horizontal net force acts on the apple to change its velocity.

What can you say about the distance traveled by the apple and Susan in the same amount of time, i.e. the time it takes the apple to land?

Student Response

The horizontal distances are the same because both Susan and the apple travel with same velocity in the same amount of time.

Current Task

Susan is riding her bike with constant velocity to school while holding an apple. On the way, she accidentally drops the apple. Assume there is no air resistance.

Where does that apple land relative to Susan along the road?

Please begin by briefly answering the above question. After briefly answering the above question, please go on to explain your answer in as much detail as you can.

Multimedia

Copyright © DeepTutor.org, 2012

Figure 1. DeepTutor Physics Problem and Dialogue History.

The screenshot of DeepTutor shows a Physics problem on the top right pane and a Dialogue history on the left pane. The Multimedia box synchronizes with the dialogue showing identified information, for example, velocity and force vectors, visually.

such descriptions of how students learn, "assessments can be designed to identify current student thinking, likely antecedent understandings, and next steps to move the student toward more sophisticated understandings" (National Research Council [2001], p. 182). This was basically a call for developing learning progressions (Corcoran, Mosher, and Rogat 2009).

LPs adopt a learner-centered view of a topic by modeling students' successful paths toward mastery as opposed to paths prescribed by domain experts following a logical decomposition of the big ideas of a domain. The logical decomposition provided by experts could be useful as a starting point that needs to be reorganized based on evidence of how students actually develop mastery of the big ideas. These actual, developmentally proven paths must be documented and guide instruction.

Learning progressions in DeepTutor. LPs are the central theme in DeepTutor around which everything else (domain modeling, assessment, and instructional tasks) is organized and aligned. This centrality of the LP can be seen in figure 2. Inside the circle in the middle of the figure, we show a snapshot of the LP. The shown LP is a partial view, for illustration purposes, of our Newtonian Physics LP.

The LP in DeepTutor is organized in a set of strands along the horizontal axis with strands more to the right signifying more sophisticated topics. For instance, the circular motion strand (rightmost column in the LP) is more complex compared to the Mass-and-Motion strand (third column). Along the vertical axis, each strand is organized more like a traditional LP (Alonzo and Steedle 2009), in levels of sophistication. Each level corresponds to a set of coherent ideas or models that students use to reason about the domain. The higher the level in the LP, the stronger the model, that is, the model explains more phenomena of the domain. We call our LP a two-dimensional or 2-D LP due to its organization in two dimensions. A typical LP, for example, Alonzo and Steedle's, is unidimensional (single-strand) showing only levels of sophistication for a major theme. The hierarchal structure of the LP levels within a strand can be accomplished by following several methodologies, such as how close a model is to the best model (the mastery model is the top level in an LP, also called the upper anchor), item difficulty, and also based on developmental and cognitive considerations. It should be noted that LP developers acknowledge that there is no unique, perfect progression or hierarchical structure of ideas but rather a family of

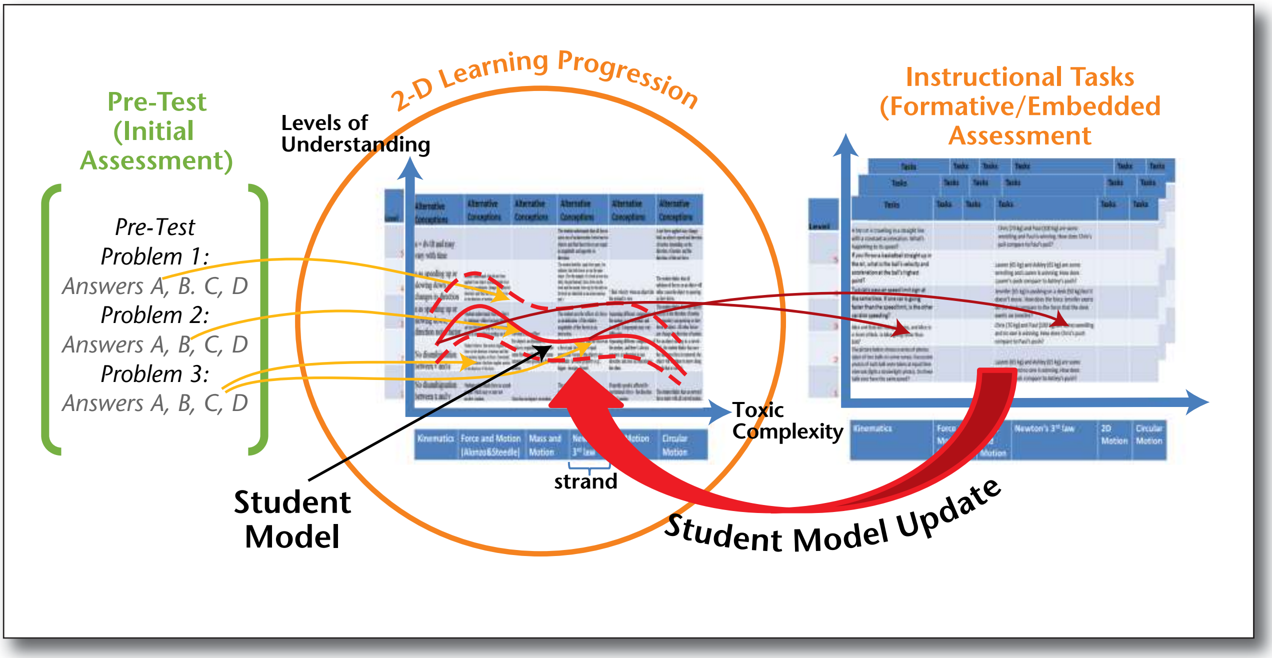


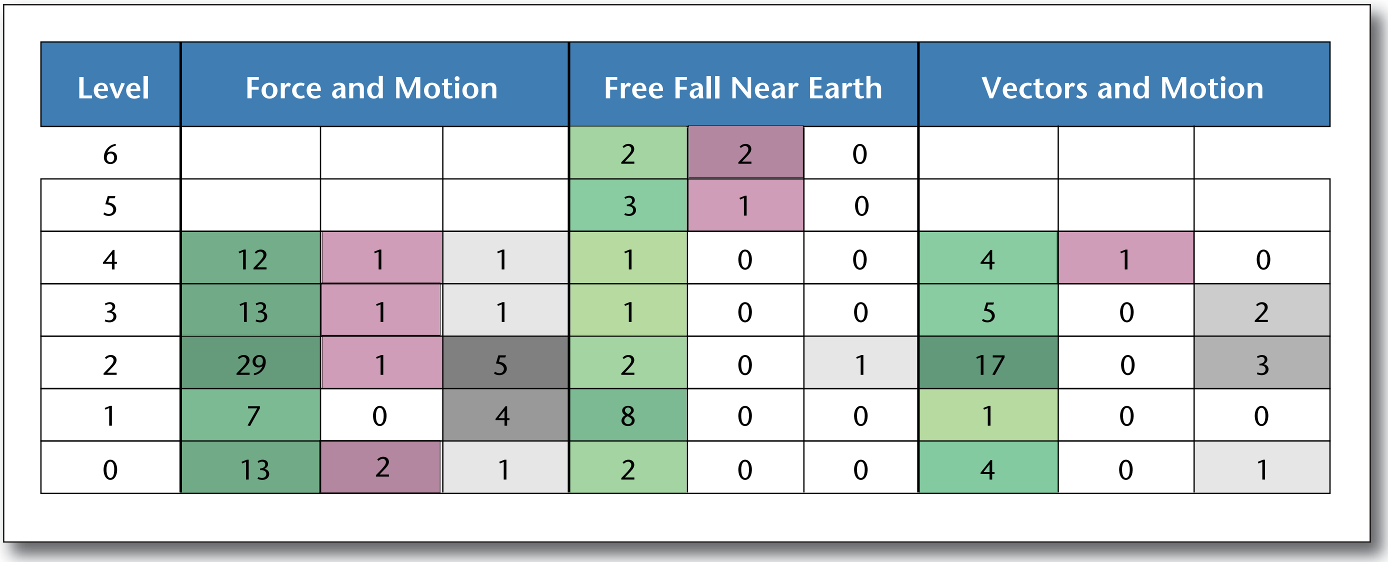
Figure 2. Alignment of Curriculum, Instruction, and Assessment in DeepTutor Using Learning Progressions.

such progressions. The goal is to document the alternatives and the most frequent progressions or levels of understanding that students experience in their journey toward the upper anchor.

Our 2-D LP is the most comprehensive and fine-grained LP compared to any other existing Physics LP. The actual LP has seven strands and up to nine levels for certain strands. It should be noted that a complex and finely grained LP is needed to drive the operations of an effective educational technology such as DeepTutor.

There is an interesting interplay among assessment, LPs, instructional tasks, and advanced tutoring strategies that is finely orchestrated by DeepTutor. The LPs are aligned with the initial assessment instrument (that is, pretest — shown on the left side of figure 2), which students must complete before they interact with the system. Based on this first summative assessment, an initial map of students' knowledge levels across all strands in the LP is generated. This corresponds to the solid, wavy line (shown in red or gray) across the middle layers of the LP in figure 2. Basically, we get a first impression of where students are with respect to the upper anchor of each LP strand. Based on recent research by others (Bao and Redish 2006; Alonzo and Steedle 2009) and our own experience, this first assessment of a student's knowledge state is just an approximation. In fact, the two wavy, dotted lines below and above the thick line indicate a range of levels that a student might be at. We are moving toward a probabilistic model (called a cloud model, inspired from the cloud model of the electron in modern physics) of assessing students' levels of understanding in which we can only assert with a certain probability at which level students are. The cloud model assumes that students can be at multiple levels in the LP, that is, have multiple models simultaneously active in their minds, some stronger than others, which they activate depending on various factors. For instance, Bao and Redish (2006) studied different student models for Newton's third law and identified the probability with which a student activates a particular model based on three features of instructional tasks that target Newton's third law.

Our cloud model for assessing students' knowledge states is work in progress as of this writing. A glimpse of it can be seen in figure 3 where we show three strands of our LP and an actual student model indicating his performance with respect to the LP levels. The leftmost column indicates the level in the LP. Each of the other columns represents an LP-strand. Each strand is further divided into three inner-columns: left — number of potential answer choices in the pretest that map to the corresponding LP level; middle — number of correct student-chosen answer choices; right — number of incorrect student-chosen answer choices. Because each multiple-choice question in the pretest has five choices of which only one is correct and because the answer choices of a question can map to different levels and strands in the LP, the relationship among the numbers shown in figure 2 is more complex. A simple pattern does exist: the number of (correct and incorrect) student answer choices should add up to the number of questions in the pretest. From the figure, we can notice that this particular student has two relatively persistent models for the force and motion topic. These two models correspond to the LP levels most chosen by the student: level 1 (corresponding



Level	Force and Motion			Free Fall Near Earth			Vectors and Motion		
6				2	2	0			
5				3	1	0			
4	12	1	1	1	0	0	4	1	0
3	13	1	1	1	0	0	5	0	2
2	29	1	5	2	0	1	17	0	3
1	7	0	4	8	0	0	1	0	0
0	13	2	1	2	0	0	4	0	1

Figure 3. The Visualization of a Student's Level of Understanding Along Three LP Strands: Force and Motion, Free Fall Near Earth, and Vectors and Motion.

to the four incorrect answer choices picked by the student) and level 2 (corresponding to the six student choices of which five are incorrect and one correct). These are weak models because they are in the mid-lower part of the force and motion strand. Based on the placement of the student in the LP, instructional tasks that are appropriate for levels 1 and 2 in the force and motion LP strand are selected for this particular student.

Tailored Learning Trajectories.

The set of instructional tasks that are selected at the outer loop should, ideally, be tailored to each individual learner because learners start interacting with an ITS at different levels of understanding, and also they learn at different paces. We call the set of instructional tasks a student is assigned a learning trajectory. It should be noted that there is a distinction between the hypothetical learning trajectory, which is first computed based on a student's pretest performance, and the actual learning trajectory, which is a dynamically updated version of the hypothetical learning trajectory based on students' actual performance while working on tasks (Battista 2011).

In DeepTutor, we select the hypothetical learning trajectory based on the alignment between the (summative and formative) assessment and the LPs, on one hand, and the LPs and instructional tasks, on the other hand. We have tasks corresponding to each cell in our 2-D LP. In fact, for each cell we have a set of tasks that enable us to implement advanced tutoring strategies to address some of the illusions of tutoring (Graesser, D'Mello, and Person 2009). For instance, consider the "illusion of mastery," the unwarranted assumption that the student has mastered much more than the student has really mastered. Indeed, one provocative finding in the tutoring literature is that there is sometimes a positive correlation between a student's knowledge of the material (based on pretest or posttest scores) and the student's likelihood of saying no rather than yes to the tutors' comprehension-gauging questions (Graesser, Person, and Magliano 1995). Thus, it is the knowledgeable tutees who tend to say "No, I don't understand." We detect the illusion of mastery in DeepTutor and activate a tutoring strategy to penetrate the potential illusion. For instance, one way to detect it is to identify students who keep saying "I understand" when asked if they understood a previously discussed idea or concept but actually have a low pretest score. The tutoring strategy triggered for such students consists of challenging them to solve a similar problem in order to demonstrate that they truly understood the concept. This is why each cell in the 2-D LP has a set of corresponding instructional tasks (see the layered tables on the right side of figure 2). Each table indicates a set of LP-aligned tasks. Tasks in corresponding cells in the different tables are equivalent. Sets of equivalent tasks for the same cell in the LP are needed such that DeepTutor can trigger as many tasks at a particular level of understanding as needed.

There are many possible learning trajectories that can be chosen for a particular student. For instance, drilling tasks that offer training on one big idea or theme covered in one LP strand will more likely help students move up the level of understanding within that strand while not making progress on ideas in other strands. For some strands, for example, Newton's third law, for which the correct answer to many problems is the same (the tasks are isomorphic to some degree), such a drilling training strategy is less effective. Students may learn the jingle ("the action and reaction forces are equal and opposite") after seeing the solution to a few problems and just recite the jingle when prompted to solve subsequent tasks, without actually developing a deep understanding of Newton's third law. Smarter sequencing of problems must be adopted as isomorphic problems lead to copying and therefore shallow learning (VanLehn 2011, Renkl 2002).

DeepTutor is still under development as of this writing, but components of the system have been validated. For instance, we have validated our hypothetical learning progressions (HLP) designed by our experts based on data collected from students in six different high school classes that form a developmental progression: Physics, Honors Physics, IB Physics 1, IB Physics 2, AP Physics C-1, and AP Physics C-2. The data was collected at one time across all these classes. That is, different students were accounted for in different classes. Ideally, a longitudinal study is needed to observe how a group of students' taking these classes progress over time. The adopted procedure is the best approximation within the constraints of collecting data in one semester.

Deeper Dialogue and Natural Language Processing

In this section, we present DeepTutor's advanced dialogue and natural language-processing algorithms. We distinguish between dialogue processing components that handle aspects of managing the dialogue and components that deeply understand students' contributions as a way to assess their level of understanding.

Dialogue Processing

The dialogue manager in DeepTutor implements conversational goals used in existing dialogue-based tutoring systems, such as coaching students to articulate expectations, correcting students' misconceptions, and attempting to answer students' questions when they are sufficiently inquisitive to ask questions. However, DeepTutor has additional conversational goals that attempt to achieve accurate grounding at each turn (that is, the system and tutor perfectly understand each other at every turn and over many turns), accurate feedback on students' contributions, error-handling (for cases when the system cannot accurately interpret what the student is saying), naturalistic dialogue, and optimized knowledge transfer. To achieve the new goals, we added components that explicitly handle dialogue moves associated with these goals to the core dialogue management module. For instance, we have a component to detect the need for establishing common ground and another to initiate and handle the dialogue moves necessary to establish common ground. Consider the scenario in which the tutoring system presents a rare or unseen word *X* to a student and the student replies with "What is *X?"* That is an indication of a request for grounding. By the same token, we can imagine a student replying "Yes" to a comprehension-gauging question ("Do you understand?") asked by the tutoring system. The system would then skeptically challenge the student with a verification statement to double-check the student's understanding. If the student stumbles, it is an indication of knowledge transfer failure and thus the system must activate a component to optimize the transfer of knowledge.

Speech Act Classification

An important component of the dialogue manager is the identification of students' intentions based on their utterances, that is, the task of speech act classification (SAC; Rus et al. [2012b]). The SAC uses a multileveled taxonomy of speech act categories with 4 major categories at the top (metacognitive, metacommunicative, question, and contribution) and 35 categories total at the second and third level. In case a student utterance is labeled as being a contribution, which is a content-rich statement, it is passed on to the semantic processing component that is described in the next section. Contributions from students can be relevant or irrelevant (for example, a content-rich statement that talks about food or friends when the topic is physics). Only relevant contributions are passed on for further semantic analysis.

Advanced Algorithms for Understanding Students' Natural Language Essays and Contributions

Algorithms are needed to interpret the meaning of students' natural language contributions at each turn in the dialogue as well as assessing the more comprehensive essay-type answers that students are required to provide immediately after being prompted to solve a problem. This section describes advances including the addition of negation handling and syntactic information as well as proposing algorithms that incorporate optimized semantic matching solutions.

Semantic similarity is the underlying principle for understanding student contributions. We assess how similar a student contribution is to an expert answer. The expert answer is deemed correct and therefore the student contribution is deemed correct if it is semantically similar to the expert answer (and incorrect otherwise). More nuanced assessments are made (for example, partially correct or partially correct and partially incorrect at the same time) but we focus here on simple binary judgments (correct versus incorrect). The alternative to the semantic similarity approach is the true or full understanding approach in which the student response is fully interpreted. The full understanding approach is intractable for real-world, sizeable applications such as ITSs because it requires vast amounts of world and domain knowledge. Domain knowledge and world knowledge are captured to some extent by semantic similarity approaches.

It should be noted that because student contributions can vary in length, from very short, for example, including just one content word such as *equal*, to a sentence or even a paragraph, methods that apply across different granularities of texts are needed. The methods presented next are generally applicable to texts of various sizes although some are more suited for a certain granularity level. For instance, when syntactic information is used in a particular method then the method cannot be applied at the word level directly.

We present next one method addressing the task of semantic similarity in the context of dialogue-based ITSs, which we explored and embedded in our SEMILAR tool (Rus et al. 2013) for exploring semantic similarity methods. Several broad categories of semantic similarity methods were investigated and are included in our SEMILAR toolkit: vectorial methods including LSA (Landauer et al. 2007, Lintean et al. 2010), probabilistic methods including Latent Dirichlet Allocation (LDA; Blei, Ng, and Jordan [2003]; Niraula et al. 2013), greedy methods, optimal methods (Rus et al. 2012a; Rus and Lintean 2012), and some others. Due to space reasons, we only discuss one of the most recent and most advanced methods we developed.

Optimal Word-to-Word and Syntactic Matching Through Quadratic Assignment

This method looks for an optimal global assignment of words in one sentence (for example, a student response) to words in the other sentence (for example, the expert answer) based on their word-to-word similarity, while simultaneously maximizing the match between the syntactic dependencies. Accounting for the syntactic dependencies among words is the primary advantage of the quadratic assignment problem (QAP; Koopmans and Beckmann [1957]) formulation versus the job-assignment formulation of the student response assessment task (Rus and Lintean 2012). We formulate first the quadratic assignment problem and then show how we model the semantic similarity task based on it.

The goal of the Koopmans-Beckmann (1957) formulation of the QAP problem, first proposed for economic activities, is to minimize the objective function QAP (see below) where matrix F describes the flow between any two facilities, matrix D indicates the distances between locations, and matrix B provides the cost of locating facilities to specific locations. F, D, and B are symmetric, nonnegative matrices.

$$\min QAP(F,D,B) = \sum_{i=1}^{n}\sum_{j=1}^{n} f_{i,j} d_{\pi(i)\pi(j)} + \sum_{i=1}^{n} b_{i,\pi(i)}$$

The $f_{(i,j)}$ term denotes the flow between facilities i and j, which are placed at locations $\pi(i)$ and $\pi(j)$, respectively. The distance between these locations is $d_{(\pi(i)\pi(j))}$. In our case, F and D describe dependencies between words within one sentence (or the other, respectively) while B captures the word-to-word similarity between words in opposite sentences. Also, we have weighted each term in the above formulation and instead of minimizing the sum we are maximizing it, resulting in the following formulation:

$$\max QAP(F,D,B) \\ = \alpha \sum_{i=1}^{n}\sum_{j=1}^{n} f_{i,j} d_{\pi(i)\pi(j)} + (1-\alpha)\sum_{i=1}^{n} b_{i,\pi(i)}$$

The $f_{(i,j)}$ term quantifies the syntactic relation between words i and j in text A, which are mapped to words $\pi(i)$ and $\pi(j)$ in text B, respectively. The distance $d_{(\pi(i)\pi(j))}$ quantifies the syntactic relation between words $\pi(i)$ and $\pi(j)$. For words i and j that have a direct dependency relation, that is, an explicit syntactic relation among two words, such as subject or direct object, the flow $f_{(i,j)}$ is set to 1 and 0 otherwise. Similarly, the distance $d_{(\pi(i)\pi(j))}$ between words $\pi(i)$ and $\pi(j)$ is set to 1 in case there is a direct dependency relation among them and 0 otherwise. We also experimented with a variant in which we enforced that the type of dependency between words i and j and the type of dependency between the corresponding words in the other text, $\pi(i)$ and $\pi(j)$, be identical. That is, we prefer matchings between words in texts A and B, respectively, that not only lead to direct dependencies among words in A and the corresponding matched words in B but those dependencies must be of the same type.

A brute force solution to the QAP problem, which would generate all possible mappings from facilities (words in a sentence) to locations (words in the other sentence), is infeasible because the solution space is too large. For example, when considering all possible pairings of words between sentence A, of size n, and sentence B of size m, where $n < m$, and we pose no limitations on the type of pairings that can be made, there are $m!/(m - n)!$ possible solutions. For sentences of average size $n = m = 20$ words, there are 2.4 □ 10^{18} possible pairings.

An efficient branch-and-bound algorithm has been developed to reduce the explored space in search for the optimal solution. This is possible by defining a bounding function that always overestimates or underestimates solutions, depending on what type of optimal solution is sought, maximum or minimum cost, respectively. In our case, a maximum cost solution is desired. Our solution yielded the best performance (accuracy = 77.6 percent) reported so far on a benchmark text-to-text similarity task at sentence level, that is, the Microsoft Research Paraphrase corpus (Dolan, Quirk, and Brockett 2004).

Designing for Affect

Another important advancement of the ITSs' inner loop focuses on techniques that target students' affective states. Although ITSs and other advanced learning technologies (ALTs) have traditionally paid less attention to affect, the tide is gradually shifting as affect-enabled (affect-aware, affect-sensitive, or affective) learning technologies are coming online. These systems are motivated by a preponderance of research that suggests that cognition and affect are inextricably coupled, and meaningful learning always encompasses some blend of the two. In this section, we briefly discuss some of the research in the growing area of affect-enabled learning technologies and closely examine one system that takes a somewhat nonconventional route toward leveraging the affect-cognition relationship to increase learning gains.

Types of Affect-Enabled Learning Technologies

One simple categorization of affect-enabled technologies broadly distinguishes between proactive and reactive systems. Proactive systems aspire to increase the incidence of affective states deemed beneficial (for example, interest, curiosity, engagement), while simultaneously decreasing the incidence of certain negative states (for example, boredom). For example, some ALTs aspire to leverage the facilitating effects of games on engagement by implementing some of the competitive and motivational features of games. Others take this a step further by carefully embedding the learning content in games that support narrativity, realism, and immersion.

Reactive systems make no notable a priori attempt to up-regulate or down-regulate positive and negative affect, respectively. Instead, they simply detect and respond to affective states as they arise. This involves fully automated detection of learner affect, which is accomplished with top-down predictive models independently or in conjunction with bottom-up sensor-driven models. Once the learner's affective state is detected with reasonable accuracy, the system then dynamically alters its pedagogical plan in response to the sensed state.

A number of reactive affect-enabled ALTs have been developed and tested. The Affective AutoTutor is a dialog-based ITS that monitors contextual cues (for example, learner response accuracy, response time, tutor feedback), body movements, and facial features to detect when a learner is confused, bored, or frustrated, and responds with motivational dialogue moves to encourage learners to persist in their learning despite these negative emotions (D'Mello et al. 2010). Forbes-Riley and Litman (2011) have recently endowed a speech-enabled dialog-based physics ITS with the ability to tailor its actions on the basis of the uncertainty and correctness of learner responses. Uncertainty is detected by fusing an acoustic-prosodic analysis of the learners' spoken responses with features extracted from the ensuing tutorial dialogue (for example, turn number). Yet another example is Gaze Tutor, a learning environment for biology that monitors eye movements to infer when learners are disengaged or zoning out and launches gaze-reactive dialogues in an attempt to reengage the learners (D'Mello et al. 2012).

Although the proactive and reactive strategies represent the major research thrusts in the area of affect-enabled ALTs, there are additional possibilities as well. One possibility is to intentionally induce certain affective states that have a somewhat counterintuitive relationship with learning. Confusion is a particularly compelling example of such a state. Though most would consider confusion to be a negative affective state, both in terms of its subjective experience (that is, most people do not like being confused) and its assumed impact on learning (that is, intuition suggests that confusion is harmful to learning), there is some correlational evidence that suggests a positive relationship between confusion and learning gains (D'Mello et al., 2013). We were intrigued by this finding, so we developed a learning environment to investigate whether there are any benefits to intentionally inducing confusion.

Case Study of an Affect-Enabled ALT: Using Confusion to Promote Clarity

The idea that some forms of confusion can be beneficial to learning is grounded in theories that highlight the benefits of impasses, cognitive conflict, and cognitive disequilibrium to learning at deeper levels of comprehension. These theories suggest that confusion signals that something is wrong with the learner's state of knowledge and can engender deeper modes of processing to the extent that the learner engages in effortful cognitive activities to resolve the confusion. Importantly, it is not the confusion itself, but the cognitive activities that accompany confusion resolution that lead to any improvement in learning gains.

A Model of Confusion Induction and Resolution

Figure 4 depicts some of the key processes of our model. ALTs, by and large, strive to promote clarity and understanding by providing hints, explanations, elaborated feedback, and other cognitive scaffolds. This might lead to illusions of clarity and understanding, but not necessarily to deep conceptual understanding. An alternate approach is to temporarily suspend clarity by interleaving contradictions, erroneous opinions, inaccurate feedback, and other confusion-inducing events. The idea is that the induced confusion will inspire learners to actively engage in deliberation, problem solving, and other forms of sense making in order to restore clarity by resolving their confusion. The learning environment can also assist by providing hints and targeted explanations when confusion resolution fails and the learner risks being hopelessly confused.

The success of this strategy will ultimately depend on the extent to which confusion has been induced and effectively resolved. Learning is not expected to be affected if the induction fails or if the induced confusion is simply ignored. Learning might also be

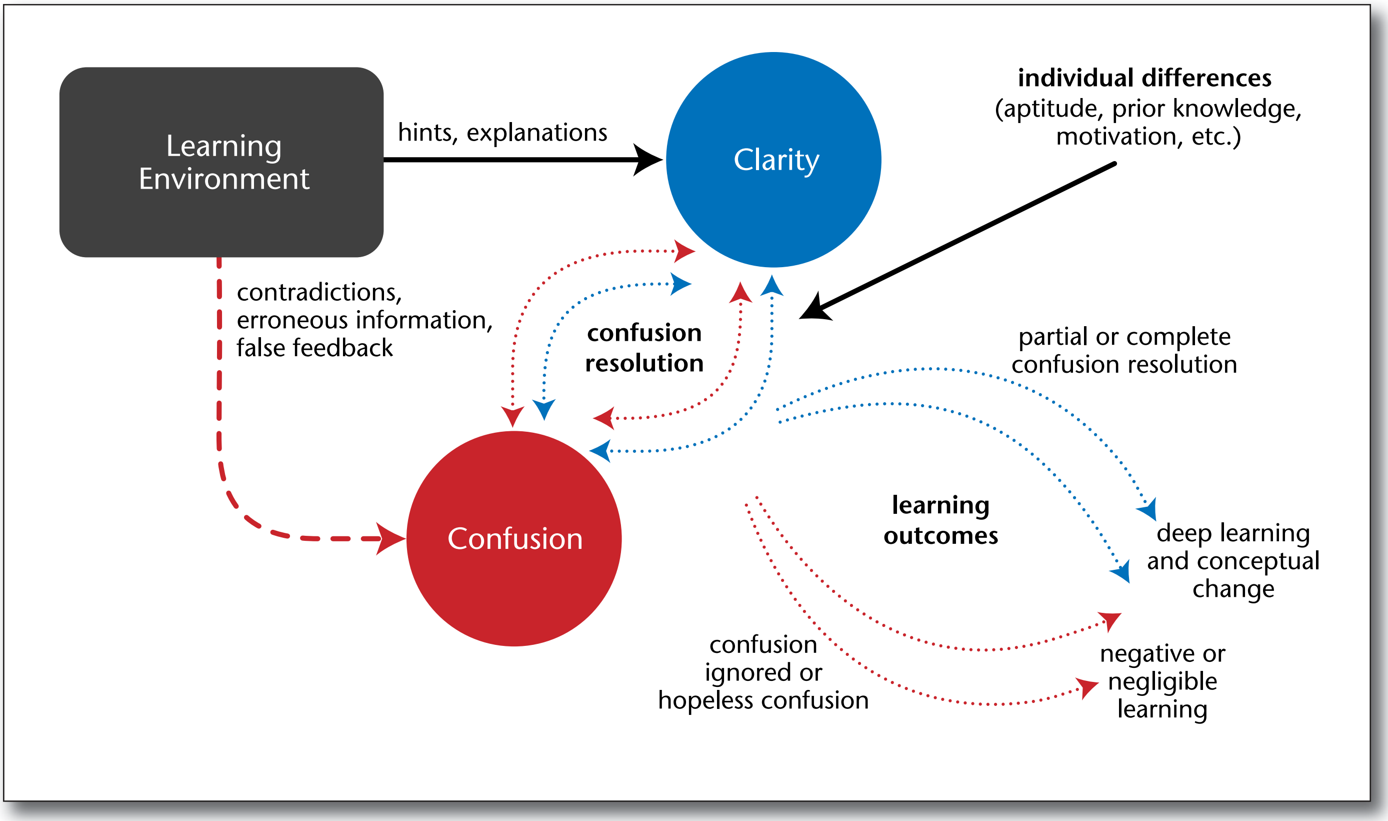

Learning Environment
hints, explanations
Clarity
individual differences
(aptitude, prior knowledge, motivation, etc.)
contradictions, erroneous information, false feedback
confusion resolution
partial or complete confusion resolution
Confusion
learning outcomes
deep learning and conceptual change
confusion ignored or hopeless confusion
negative or negligible learning

Figure 4. Model of Confusion Induction and Resolution.

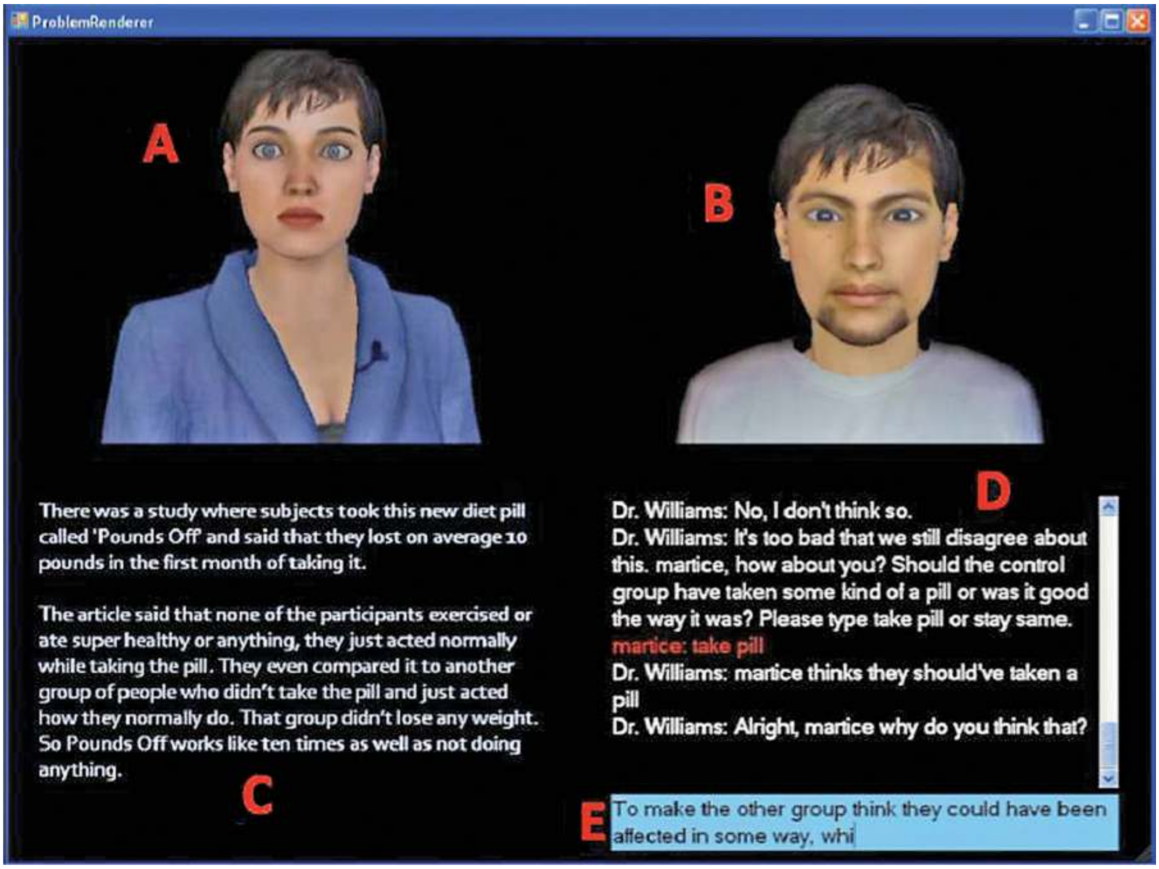
negatively affected if the induced confusion is too severe, the pedagogical scaffolds are insufficient, and learners are hopelessly confused. However, some form of deep learning is expected if learners effortfully engage in confusion-resolution processes, even if they do not fully resolve their confusion. Learners might also experience a form of conceptual change if problematic misconceptions are identified and corrected during confusion resolution. This is, of course, the sine qua non of learning.

Testing the Model

We tested the model by developing a learning environment that strategically induced confusion at critical points during the learning of scientific research methods, such as learning about the importance of random assignment to substantiate causal claims. The multimedia learning environment, modeled after the educational game Operation ARA (Halpern et al. 2012), taught research method concepts by presenting example cases of studies (including the research design, participants, methods, results, and conclusions) that were frequently flawed. Learners were instructed to evaluate the merits of the studies and point out problems over multiple trials. The critiques were accomplished by holding multiturn conversations in natural language with embodied conversational agent(s) and the human learner.

One rendition of the system (contradiction-only version) included a tutor agent who led the tutorial lessons and served as an expert on scientific inquiry and a peer agent who simulated a peer of the human learner. Confusion was induced by having the animated agents contradicting each other by occasionally disagreeing on ideas and voicing inaccurate information, and asking the human learner to intervene and decide which opinion had the most scientific merit. There was also a delayed-contradiction version in which the agents initially agreed on a concept but abruptly contradicted each other as the conversations progressed. A third version of the system (false-feedback version) implemented dialogues between the tutor agent and the human learner (there was no peer agent). This version used a false-feedback manipulation to induce confusion in which the tutor provided positive feedback to incorrect learner responses and negative feedback to correct responses.

The interface for contradiction-only and delayed-contradiction versions shown in figure 5 consisted of the tutor agent (*A*), the peer agent (*B*), a description of the research case study (*C*), a text-transcript of the dialogue history (*D*), and a text box for learners to enter and submit their responses (*E*). The agents delivered the content of their utterances through



There was a study where subjects took this new diet pill called 'Pounds Off' and said that they lost on average 10 pounds in the first month of taking it.

The article said that none of the participants exercised or ate super healthy or anything, they just acted normally while taking the pill. They even compared it to another group of people who didn't take the pill and just acted how they normally do. That group didn't lose any weight. So Pounds Off works like ten times as well as not doing anything.

Dr. Williams: No, I don't think so.
Dr. Williams: It's too bad that we still disagree about this. martice, how about you? Should the control group have taken some kind of a pill or was it good the way it was? Please type take pill or stay same.
martice: take pill
Dr. Williams: martice thinks they should've taken a pill
Dr. Williams: Alright, martice why do you think that?

To make the other group think they could have been affected in some way, whi

Figure 5. Interface for Contradiction-Only and Delayed-Contradiction Learning Sessions.

synthesized speech while the human learner typed his or her responses. Text transcripts of the trialogues were stored in log files for offline analysis.

The different versions also varied the amount of help provided to learners to resolve their confusion. The delayed-contradiction and the false-feedback versions provided learners with a short explanatory text to read after the confusion inductions. The contradiction-only version provided no explicit scaffold, with the exception that all misleading information (contradictions and false feedback) was corrected at the end of the discussion for each case study (this was the same in all versions).

We conducted four experiments (total $N = 474$) in which college students learned between four and eight research method concepts by discussing the scientific merits of sample research studies with the animated agent(s). Experiments 1 and 2 used the contradiction-only version (D'Mello et al., in press), Experiment 3 used the delayed-contradiction plus explanatory text version (D'Mello et al., in press), and Experiment 4 used the false-feedback plus explanatory text version (Lehman et al., unpublished). All four experiments used a within-subjects design in that learners analyzed some case studies with the confusion-induction methods enabled and other studies with these methods disabled.

Confusion was tracked through offline cued-recall procedures (experiments 1 and 2), online self-reports (experiments 3 and 4), and learner behaviors (for example, response times after confusion induction, accuracy on probing questions immediately following the manipulations — all experiments). Learning was measured through multiple-choice retention tests (all experiments) and a transfer task involving detection of flaws in new case studies (experiments 3 and 4).

The results of all four experiments generally indicated that the manipulations had the desired effect of inducing confusion. Importantly, learning was not affected by the manipulations alone or when the manipulations failed to induce confusion. However, performance on the learning measures was substantially higher in the experimental conditions compared to the control conditions when the manipulations were successful in inducing confusion.

The results of these four experiments are significant because they constitute some of the first experimental evidence on the moderating effect of confusion on learning. The most obvious implication is that there might be some benefits for ALTs that intentionally perplex learners. Of course, this is only if learners have the requisite knowledge and skills to resolve the confusion and the learning environment provides appropriate scaffolds to help with the confusion-resolution process. But, when done right, confusion is one road to clarity.

Concluding Remarks

Much has changed in the last decade since Graesser and colleagues' (2001) article. We now know that conversational ITSs are more effective than noninteractive text controls, their effectiveness depends on the extent of the alignment between the content and students' prior knowledge, they are as effective as nonconversational ITSs, and they rival or outperform novice human tutors on comparable content (Olney et al. 2012; VanLehn et al. 2007; VanLehn 2011). We also know that the content of the tutorial dialogues is the primary driver of learning gains, at least when compared to other aspects of the interaction, such as whether the tutors or students communicate through speech or text or whether animated pedagogical agents are present or absent.

Though several of the earlier predictions pertaining to the hypothesized effectiveness of conversational ITSs have generally been supported, two counterintuitive findings, one positive and one negative, stand out. On the positive front, a recent meta-analysis by VanLehn (2011) revealed that the so called "expert" human tutors rarely achieve the much touted two sigma effect noted by Bloom (1984) and that conversational ITSs are equally effective as these human tutors. This positive finding is tempered by the fact that the effectiveness of conversational tutors plateaus at about one sigma (approximately a letter grade), a point at which increased interactivity has little or no additional impact in increasing learning.

This article has focused on some of our recent research on next-generation conversational agents that aspire to ascend the interactional plateau into the land of two sigma effects. We have described DeepTutor, a conversational ITS that implements learning progressions and deeper natural language understanding. We have also discussed affect-aware ITSs that take the student-tutor interaction to the next level by modeling both the affective and the cognitive aspects of learning.

Our hope for the next decade is a world in which these and other conversational ITSs autonomously help hundreds of thousands of students develop content mastery, learning strategies, critical thinking, writing proficiency, and other 21st century skills in a manner that effectively integrates cognition, motivation, and emotion. This is a tall order indeed, but the task of educating the next generation of students is not for the faint of heart.

Acknowledgements

The authors would like to thank members of DeepTutor project,[4] the Tutoring Research Group,[5] and the Emotive Computing Group[6] at the University of Memphis. This research was supported by the Institute for Education Sciences (IES) under award R305A100875 to Vasile Rus and National Science Foundation (NSF) under awards ITR 0325428, HCC 0834847, and DRL 1235958 to Arthur C. Graesser and Sidney D'Mello. Any opinions, findings and conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the IES or NSF.

Notes

1. See www.deeptutor.org.
2. See www.authortutor.org.
3. More information about the SEMILAR toolkit is available at www.semanticssimilarity.org.
4. See www.deeptutor.org.
5. See www.autotutor.org.
6. See emotion.autotutor.org.

References

Alonzo, A. C., and Steedle, J. T. 2009. Developing and Assessing a Force and Motion Learning Progression. *Science Education* 93(3): 389–421.

Bao, L., and Redish, E. F. 2006. Model Analysis: Assessing the Dynamics of Student Learning. *Physical Review Special Topics Physics Education Research* 2(1).

Battista, M. T. 2011. Conceptualizations and Issues Related to Learning Progressions, Learning Trajectories, and Levels of Sophistication. *The Mathematics Enthusiast* 8(3): 507–569.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3: 993–1022.

Bloom, B. 1984. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher* 13(6): 4–16.

Chi, M. T. H.; Siler, S. A.; and Jeong, H. 2004. Can Tutors Monitor Students' Understanding Accurately? *Cognition and Instruction* 22(3): 363–387.

Corcoran, T.; Mosher, F. A.; and Rogat, A. 2009. Learning Progressions in Science: An Evidence-Based Approach to Reform. Consortium for Policy Research in Education Report #Rr-63. Philadelphia, PA: Consortium for Policy Research in Education.

D'Mello, S.; Lehman, B.; Sullins, J.; Daigle, R.; Combs, R.; Vogt, K.; Perkins, L.; and Graesser, A. 2010. A Time for Emoting: When Affect-Sensitivity Is and Isn't Effective at Promoting Deep Learning. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, ed. J. Kay and V. Aleven, 245–254. Berlin: Springer.

D'Mello, S.; Olney, A.; Williams, C.; and Hays, P. 2012. Gaze Tutor: A Gaze-Reactive Intelligent Tutoring System. *International Journal of Human-Computer Studies* 70(5): 377–398.

D'Mello, S., Lehman, S., Pekrun, R., and Graesser, A. 2013. Confusion Can Be Beneficial for Learning. Learning and Instruction.

Dolan, W. B.; Quirk, C.; and Brockett, C. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of the 20th International Conference on Computational Linguistics,* Geneva, Switzerland.

Dynarski, M.; Agodini, R.; Heaviside, S.; Novak, T.; Carey, N.; Campuzano, L. 2007. Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Evens, M., and Michael, J. 2005. *One-on-One Tutoring by Humans and Machines*. Mahwah, NJ: Lawrence Erlbaum Associates.

Forbes-Riley, K., and Litman, D. J. 2011. Benefits and Challenges of Real-Time Uncertainty Detection and Adaptation in a Spoken Dialogue Computer Tutor. *Speech Communication* 53(9–10): 1115–1136.

Graesser, A. C.; D'Mello, S.; and Person, N. K. 2009. Meta Knowledge in Tutoring. In

Handbook of Metacognition in Education, ed. D. Hacker, J. Dunlosky, and A. C. Graesser. Mahwah, NJ: Taylor and Francis.

Graesser, A.; Moreno, K.; Marineau, J.; Adcock, A.; Olney, A.; and Person, N. 2003. AutoTutor Improves Deep Learning of Computer Literacy: Is It the Dialog or the Talking Head? In *Proceedings of the 11th International Conference on Artificial Intelligence in Education*, ed. U. Hoppe, F. Verdejo, and J. Kay, 7–54. Amsterdam: IOS Press.

Graesser, A. C.; Person, N. K.; and Magliano, J. P. 1995. Collaborative Dialogue Patterns in Naturalistic One-to-One Tutoring. *Applied Cognitive Psychology* 9(6): 495–522.

Graesser, A. C.; Rus., V.; D'Mello, S.; and Jackson, G. T. 2008. AutoTutor: Learning Through Natural Language Dialogue That Adapts to the Cognitive and Affective States of the Learner. In *Current Perspectives on Cognition, Learning, and Instruction: Recent Innovations in Educational Technology That Facilitate Student Learning*, ed. D. H. Robinson and G. Schraw, 95–125. Charlotte, NC: Information Age Publishing.

Graesser, A. C.; VanLehn, K.; Rose, C. P.; Jordan, P.; and Harter, D. 2001. Intelligent Tutoring Systems with Conversational Dialogue. *AI Magazine* 22(4): 39–41.

Halpern, D. F.; Millis, K.; Graesser, A. C.; Butler, H.; Forsyth, C.; and Cai, Z. 2012. Operation ARA: A Computerized Learning Game That Teaches Critical Thinking and Scientific Reasoning. *Thinking Skills and Creativity* 7(93–100).

Koopmans, T. C., and Beckmann, M. 1957. Assignment Problems and the Location of Economic Activities. *Econometrica* 25(1): 53–76.

Landauer, T.; McNamara, D. S.; Dennis, S.; and Kintsch, W., eds. 2007. *Handbook on Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Lehman, B.; D'Mello, S.; and Graesser, A. Unpublished. False Feedback, Confusion, and Learning.

Mohan, L.; Chen, J.; and Anderson, W. A. 2009. Developing a Multi-Year Learning Progression for Carbon Cycling in Socioecological Systems. *Journal of Research in Science Teaching* 46(6): 675–698.

National Research Council. 2001. *Knowing What Students Know: The Science and Design of Educational Assessment,* ed. J. W. Pellegrino, N. Chudowsky, and R. Glaser. Washington, DC: National Academy Press.

Niraula, N.; Banjade, R.; Stefanescu, D.; and Rus, V. 2013. Experiments with Semantic Similarity Measures Based on LDA and LSA. In *Proceedings of the First International Conference on Statistical Language and Speech Processing,* Lecture Notes in Computer Science, 7978. Berlin: Springer.

Olney, A.; D'Mello, S.; Person, N.; Cade, W.; Hays, P.; Williams, C.; Lehman, B.; and Graesser, A. C. 2012. Guru: A Computer Tutor That Models Expert Human Tutors. In *Proceedings of the 11th International Conference on Intelligent Tutoring Systems,* ed. S. Cerri, W. Clancey, G. Papadourakis, and K. Panourgia, 256–261. Berlin: Springer.

Renkl, A. 2002. Worked-Out Examples: Instructional Explanations Support Learning by Self-Explanations. *Learning and Instruction* 12(5): 529–556.

Rus, V., and Graesser, A. C. 2006. Deeper Natural Language Processing for Evaluating Student Answers in Intelligent Tutoring Systems. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence.* Menlo Park, CA: AAAI.

Rus, V., and Lintean, M. 2012. A Comparison of Greedy and Optimal Assessment of Natural Language Student Input Using Word-to-Word Similarity Metrics. Paper presented at the Seventh Workshop on Innovative Use of Natural Language Processing for Building Educational Applications, Montreal, Canada, June 7–8.

Rus, V.; Lintean, M.; Banjade, R.; Niraula, N.; and Stefanescu, D. 2013. SEMILAR: The Semantic Similarity Toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics.

Rus, V.; Lintean, M.; Moldovan, C.; Baggett, W.; Niraula, N.; Morgan, B. 2012a. The Similar Corpus: A Resource to Foster the Qualitative Understanding of Semantic Similarity of Texts. In *Semantic Relations II: Enhancing Resources and Applications, Proceedings of the 8th Language Resources and Evaluation Conference.* Paris: European Language Resources Association.

Rus, V.; Moldovan, C.; Graesser, A.; and Niraula, N. 2012b. Automatic Discovery of Speech Act Categories in Educational Games. Paper presented at the 5th International Conference on Educational Data Mining, Chania, Greece 19–21 June.

VanLehn, K. 2006. The Behavior of Tutoring Systems. *International Journal of Artificial Intelligence in Education.* 16(3): 227–265.

VanLehn, K.; Graesser, A. C.; Jackson, G. T.; Jordan, P.; Olney, A.; and Rose, C. P. 2007. When Are Tutorial Dialogues More Effective Than Reading? *Cognitive Science* 31(1): 3–62.

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

Vasile Rus is an associate professor of computer science in the Institute for Intelligent Systems at the University of Memphis. His current research focuses on building conversational intelligent tutoring systems. He is the principal investigator of the DeepTutor project funded by the Institute for Education Sciences. Rus was involved in various roles in the development of other intelligent tutoring systems including MetaTutor, W-Pal, iSTART, and AutoMentor. Rus's other research interests are knowledge representation, machine learning, text mining, and software engineering.

Sidney D'Mello is an assistant professor holding joint appointments in the Departments of Computer Science and Psychology at the University of Notre Dame. His interests include emotional processing, affective computing, artificial intelligence in education, human-computer interaction, speech recognition and natural language understanding, and computational models of human cognition. He is an associate editor for *IEEE Transactions on Affective Computing* and serves as an advisory editor for the *Journal of Educational Psychology.*

Xiangen Hu is a professor in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis and visiting professor at Central China Normal University. Hu's primary research areas include mathematical psychology, research design and statistics, and cognitive psychology. More specific research interests include general processing tree (GPT) models, categorical data analysis, knowledge representation, computerized tutoring, and advanced distributed learning.

Art Graesser is a professor in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis and is an Honorary Research Fellow at the University of Oxford. His primary research interests are in cognitive science, discourse processing, emotion, tutoring, and the learning sciences. He and his colleagues have developed learning technologies with animated conversational agents (such as AutoTutor and Operation ARA) and automated analyses of texts at multiple levels (such as Coh-Metrix, and Question Understanding AID — QUAID).