AutoTutor and Affective AutoTutor: Learning by Talking with Cognitively and Emotionally Intelligent Computers that Talk Back

SIDNEY D'MELLO, University of Notre Dame
ART GRAESSER, University of Memphis

We present AutoTutor and Affective AutoTutor as examples of innovative 21st century interactive intelligent systems that promote learning and engagement. AutoTutor is an intelligent tutoring system that helps students compose explanations of difficult concepts in Newtonian physics, computer literacy, and critical thinking by interacting with them in natural language with adaptive dialogue moves similar to those of human tutors. AutoTutor constructs a cognitive model of students' knowledge levels by analyzing the text of their typed or spoken responses to its questions. The model is used to dynamically tailor the interaction toward individual students' zones of proximal development. Affective AutoTutor takes the individualized instruction and human-like interactivity to a new level by automatically detecting and responding to students' emotional states in addition to their cognitive states. Over 20 controlled experiments comparing AutoTutor with ecological and experimental controls such reading a textbook have consistently yielded learning improvements of approximately one letter grade after brief 30-60 minute interactions. Furthermore, Affective AutoTutor shows even more dramatic improvements in learning than the original AutoTutor system, particularly for struggling students with low domain knowledge. In addition to providing a detailed description of the implementation and evaluation of AutoTutor and Affective AutoTutor, we also discuss new and exciting technologies motivated by AutoTutor such as AutoTutor-Lite, Operation ARIES, GuruTutor, DeepTutor, MetaTutor, and AutoMentor. We conclude this paper with our vision for future work on interactive and engaging intelligent tutoring systems.

Categories and Subject Descriptors: **K.3.1 [Computers and Education]**: Computer Uses in Education

General Terms: Intelligent Tutoring Systems, Affective Computing, Natural Language Understanding, User Modeling

Additional Key Words and Phrases: constructivism, deep learning, affect detection and synthesis, student modeling

ACM Reference Format:

D'Mello, S. K., Graesser, A. C.. 2012. AutoTutor and Affective AutoTutor: Learning by Talking with a Cognitively and Emotionally Intelligent Computer that Talks Back. ACM Trans. Interactive Intelligent. Syst. X, X, Article XX (Month 2011), X pages.

DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

About a decade ago, the idea of a student learning difficult technical content, such as Newtonian physics and computer operating systems, by typing or speaking to a

---

The research on was supported by the National Science Foundation (SBR 9720314, REC 0106965, REC 0126265, ITR 0325428, REESE 0633918, NSF-0834847, DRK-12-0918409, DRL-1108845), the Institute of Education Sciences (R305H050169, R305B070349, R305A080589, R305A080594), and the Department of Defense Multidisciplinary University Research Initiative (MURI) administered by ONR under grant N00014-00-1-0600. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, or DoD.

Author's addresses: Sidney K. D'Mello, Departments of Computer Science and Psychology, U. of Notre Dame; Art C. Graesser, Department of Psychology and Institute for Intelligent Systems, U. of Memphis.

Permission to make digital or hardcopies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credits permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

@2010 ACM 1539-9087/2010/03-ART39 $10.00

DOI10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

computer in natural language would be considered by most to be a mere fantasy. A computer that gauges the student's level of knowledge by asking probing questions, analyzes the student's responses to those questions, proactively identifies and corrects misconceptions, and even responds to the student's own questions, gripes, and comments, would also have been considered to be no more than a seductive vision. Similarly, a decade ago, most would consider the idea that a computer could someday sense when a student is bored or frustrated and dynamically change its strategies to help the student conquer these negative emotions to be plainly absurd. Therefore, the fact that a computer system with these and many more capabilities has, in fact, been developed is an example of progress made over the last decade. In this paper, we present an overview of the design, implementation, and evaluation of two such systems called AutoTutor and Affective AutoTutor as examples of 21st century innovation in the field of Intelligent Interactive Systems.

AutoTutor is an Intelligent Tutoring System (ITS) that helps students learn complex technical content in Newtonian physics, computer literacy, and critical thinking by (a) holding a conversation in natural language, (b) simulating the pedagogical and motivational strategies of human tutors, (c) modeling students' cognitive states, (d) using its student model to dynamically tailor the interaction to individual students, (e) answering students' questions, (f) identifying and correcting misconceptions, and (g) keeping students engaged with images, animations, and simulations. In addition to these capabilities, Affective AutoTutor adds affect-sensitive capabilities by (h) detecting students' affective states by monitoring facial features, body language, and conversational cues, (i) regulating negative affective states such as frustration and boredom, and (j) synthesizing emotions via the content of its verbal responses, speech intonation, and facial expressions of an animated pedagogical agent. Much like a gifted human tutor, AutoTutor and Affective AutoTutor attempt to keep the student balanced between the extremes of boredom and bewilderment by subtly modulating the pace, direction, and complexity of the learning task.

The design of AutoTutor was inspired by explanation-based constructivist theories of learning [Aleven and Koedinger 2002; Bransford et al. 1991; Chi et al. 1994; Piaget 1952; Rogoff 1990; VanLehn et al. 1992; Vygotsky 1978] and by previous empirical research that has documented the collaborative constructive activities that routinely occur during human tutoring [Chi et al. 2001; D'Mello et al. 2010a; D'Mello et al. 2010e; Fox 1993; Graesser and Person 1994; Graesser et al. 1995; Moore 1995; Shah et al. 2002; VanLehn et al. 2007]. Constructivist approaches have shaped the standards for curriculum and instruction in the United States during the last decade. According to this approach, students need to actively construct coherent, explanation-based meanings and knowledge by interacting with the world and other people. Simply put, students learn by telling and doing. Learning environments should stimulate active construction of knowledge and provide feedback and explanations on these constructions rather than being mere information delivery systems. AutoTutor adheres to these constructivist principles because it was designed to simulate the dialogue moves of human tutors who coach students in constructing explanations to difficult problems.

Although the principles underlying AutoTutor strongly adhere to the cognitive elements of constructivism, it is important to note that constructivism is not entirely limited to cognition, discourse, action, and the environment because emotions (affective states) are inextricably bound to the learning process [Boekaerts 2007; Calvo and D'Mello 2011; D'Mello and Graesser 2012; Pekrun and Stephens 2012; Schutz and Pekrun 2007]. An agile learning environment that is sensitive to a student's affective states presumably enriches learning, particularly when learning is accompanied by confusion, frustration, boredom, interest, excitement, and insight.

Therefore, in addition to modeling and responding to students' cognitive states, Affective AutoTutor also detects and helps regulate negative emotional states such as boredom and frustration in order to increase engagement, task persistence, and learning.

AutoTutor and Affective AutoTutor fill an important societal need because it is widely acknowledged that the one-size-fits-all approach of most classrooms is not conducive to learning at deeper levels of comprehension. Information transmission via lecturing, which is a major classroom activity, typically fosters factual and rule-based thinking (e.g., memorizing facts and definitions), but rarely facilitates model-based reasoning and deep thinking (e.g., problem solving, analyzing causal relationships, making bridging inferences). Therefore, students turn to one-on-one human tutoring when they are having difficulty in their STEM (Science Technology Engineering and Mathematics) courses. One-on-one human tutoring does have a payoff because there is considerable empirical evidence showing that human tutoring is extremely effective when compared to typical classroom environments [Bloom 1984; Cohen et al. 1982; Corbett 2001; Fletcher 2003]. However, the cost associated with providing each student with a human tutor makes the adoption of widespread tutoring programs unfeasible, and as a consequence, many students are left behind. AutoTutor and Affective AutoTutor provide a technological solution to this problem by simulating the pedagogical and motivational aspects of human tutors in a scalable and cost-effective way.

We consider AutoTutor and Affective AutoTutor to be systems that exemplify the innovative research in Interactive Intelligent Systems that have emerged over the last decade for the reasons elaborated below:

- AutoTutor and Affective AutoTutor are unique because they were designed to closely model the pedagogical styles, dialogue patterns, language, and gestures of human tutors [Graesser et al. 1999]. They are also one of the few ITSs that help learning by engaging the students in natural language dialogues that closely mirror human-human tutorial dialogues.
- They are complex intelligent systems encompassing cutting-edge research in computational linguistics, discourse processing, affective computing, machine learning, embodied conversational agents, cognitive science, and the learning sciences [Graesser et al. 2005].
- AutoTutor has been tested on over a thousand students and produces learning gains of approximately one letter grade. AutoTutor's learning gains outperform novice human tutors and almost reach the bar of expert human tutors [VanLehn, Graesser, Jackson, Jordan, Olney and Rose 2007].
- Affective AutoTutor is one of the few fully-automated ITSs that detects and responds to students' affective states in addition to their cognitive states [D'Mello and Graesser 2010]. It was the first system that demonstrated the positive effects of affect sensitivity on deep learning [D'Mello et al. 2010d].
- These systems were designed by an interdisciplinary research team spanning Computer Science, Engineering, Psychology, Cognitive Science, Linguistics, Physics, and Education. Such interdisciplinary endeavors are essential for developing innovative artifacts ranging from entertainment to educational games.
- The AutoTutor and Affective AutoTutor research program has contributed to both the science of how people learn as well as towards engineering solutions to increase learning.
- These systems have inspired a number of next-generation systems that: (a) are scalable and can be rapidly deployed on web-based platforms (AutoTutor-Lite), (b) help students regulate cognitive and metacognitive processes during learning (MetaTutor), (c) process student responses at deeper levels of

analysis (DeepTutor), (d) implement alternate pedagogical strategies such as reciprocal teaching and vicarious learning in game-based environments (Operation ARIES), (e) provide automated mentorship during learning with serious game-based environments (AutoMentor), and (f) model the strategies and dialogue patterns of expert human tutors (Guru Tutor).

The remainder of the paper is organized as follows. Section 2 provides a brief overview of the history of ITSs, discusses some of the successful ITS systems, and summarizes recent research on the role of emotions during learning with ITSs. Section 3 provides a sufficiently detailed (but not comprehensive) description of AutoTutor with an emphasis on its dialogue structure, student model, interface, and architecture. Section 4 provides a synthesis of studies on evaluations of AutoTutor along dimensions such as conversational smoothness, ability to understand the student, accuracy of its student model, and its efficacy in promoting learning gains. While Sections 3 and 4 focus on the cognitive aspects of AutoTutor, Sections 5 and 6 describe affect-sensitive versions of AutoTutor. In Section 5, we describe how Affective AutoTutor detects and responds to student emotions, while Section 6 evaluates the accuracy of automated affect detection and discusses experiments that compare affective vs. non-affective versions of AutoTutor. Section 0 discusses some of the other versions of AutoTutor that were not emphasized in this paper as well as novel learning environments that are based on the AutoTutor technology. Finally, Section 8 concludes the paper with a discussion on some of the issues arising from the design of ITSs.

2. BACKGROUND AND RELATED WORK

2.1 From Computer-based Training (CBT) to Intelligent Tutoring Systems (ITSs)

The Intelligent Tutoring Systems enterprise was launched in the late 1970s and christened with the edited volume aptly entitled *Intelligent Tutoring Systems* [Sleeman and Brown 1982]. The goal was to develop computerized learning environments that had powerful intelligent algorithms that would optimally adapt to the student and formulate computer moves that optimized learning for the individual students.

ITSs were viewed as a generation beyond computer-based training (CBT). A prototypical CBT system involves mastery learning, such that the student (a) studies material presented in a lesson, (b) gets tested with a multiple choice test or another objective test, (c) gets feedback on the test performance, (d) re-studies the material if the performance is below threshold, and (e) progresses to a new topic if performance exceeds the threshold. The order of topics presented and tested can follow different pedagogical models that range in complexity from ordering on prerequisites [Gagne 1985] to knowledge space models and Bayesian models that attempt to fill learning deficits and correct misconceptions [Doignon and Falmagne 1998] and to other models that allow dynamic sequencing and navigation [O'Neil and Perez 2006]. Meta-analyses show effect sizes of 0.39 sigma[1] compared to classrooms [Dodds and Fletcher 2004].

These CBT systems are an important class of learning environments that can serve as tutors. However, the next generation of ITSs went a giant step further that enhanced the adaptability, grain-size, and power of computerized learning

---

[1] An effect-size measures the strength of a relationship between two variables. Cohen's d is a common measure of the effect in standard deviation units (sigma) between two samples with means M_1 and M_2 and standard deviations s_1 and s_2 [Cohen 1992]. According to Cohen (1992) effect sizes approximately equal to .3, .5, and .8 represent small, medium, and large effects, respectively. $d = (M_1 - M_2)/\sqrt{(s_1^2 + s_2^2)/2}$.

environments [Woolf 2009]. The processes of tracking knowledge (called user modeling) and adaptively responding to the student incorporates computational models in artificial intelligence and cognitive science, such as production systems, case-based reasoning, Bayesian networks, theorem proving, constraint satisfaction algorithms, and educational data mining [Anderson et al. 2005; Baker and Yacef 2009; Corbett and Anderson 1994; Romero and Ventura 2007; Woolf 2009].

The ITSs that have been successfully implemented and tested have produced learning gains with an average effect size of one sigma, which is roughly equivalent to one letter grade [Corbett 2001; VanLehn, Graesser, Jackson, Jordan, Olney and Rose 2007]. When compared to classroom instruction and other naturalistic controls, the 1.0 effect sizes obtained by ITSs is superior to the 0.39 effect for computer-based training, 0.50 for multimedia, and 0.40 effect obtained by novice human tutors [Cohen, Kulik and Kulik 1982; Corbett 2001; Dodds and Fletcher 2004; Wisher and Fletcher 2004]. It is, however, less than the 2 sigma effect obtained by expert tutors for mathematics in naturalistic contexts [Bloom 1984]. The naturalistic setting is important because ITSs and accomplished tutors have produced equivalent learning gains when face-to-face communication is replaced with computer-mediated communication [VanLehn, Graesser, Jackson, Jordan, Olney and Rose 2007]. Indeed, ITSs are highly effective in helping students learn.

2.2 Examples of Intelligent Tutoring Systems that Monitor Cognitive States

Successful systems have been developed for mathematically well-formed topics, including algebra, geometry, programming languages, i.e., the *Cognitive Tutors*, [Aleven et al. 2006; Anderson et al. 1995; Corbett 2002; Koedinger and Corbett 2006], physics such as *AutoTutor*, *Andes*, *Atlas*, and *Why/Atlas* [VanLehn, Graesser, Jackson, Jordan, Olney and Rose 2007; VanLehn et al. 2005], *electronics* [Lesgold et al. 1992], and *information technology* [Mitrovic et al. 2007].

Of particular interest is a subset of these ITSs that implement natural language dialogue that is comparable to the conversations that occur in human tutoring. It appears that there are two different mechanisms that potentially explain the effectiveness of one-on-one tutoring [Corbett et al. 1999; Graesser, Person and Magliano 1995]. The first is the sophisticated tutoring strategies that have been identified in the ITS literature [Psotka et al. 1988; Sleeman and Brown 1982; Woolf 2009]. The second is the dialogue patterns and natural language that help human tutors scaffold the student to new levels of mastery [Chi et al. 2008; Graesser, Person and Magliano 1995]. According to Graesser et al. [1995], there is something about discourse and natural language (as opposed to sophisticated pedagogical strategies) that to some extent explains the effectiveness of novice human tutors. They arrive at this conclusion because most novice human tutors are effective, but they use very few (if any) sophisticated pedagogical strategies. Perhaps a combination of sophisticated tutoring strategies and conversational patterns will produce the ideal tutoring system.

Some of the successful dialogue based ITSs include AutoTutor [Graesser et al. 2004; VanLehn, Graesser, Jackson, Jordan, Olney and Rose 2007], why-Atlas [Graesser et al. 2001b; VanLehn et al. 2002], ITSPOKE [Litman and Silliman 2004], CIRCSIM-Tutor [Shah, Evens, Michael and Rovick 2002], DC-Trains [Pon-Barry et al. 2004], Mission Rehearsal [Swartout et al. 2006], Tactical Language and Culture System [Johnson and Valente 2008]. These different computer tutors vary in the extent to which they simulate human dialogue mechanisms, but all of them attempt to comprehend natural language, formulate adaptive responses, and implement pedagogical strategies to help students learn. The present focus on AutoTutor represents an example of one dialogue-based ITS.

2.3 Beyond Cold Cognition: The Emergence of Affect-Sensitive ITSs

The late 1990s and the early 2000s witnessed an exciting infusion of ITSs that implemented sophisticated tutoring strategies such as error identification and correction, building on prerequisites, frontier learning (expanding on what the student already knows), student modeling (inferring what the student knows and using that information to guide tutoring), building coherent explanations, and natural language dialogues [Aleven and Koedinger 2002; Anderson, Douglass and Qin 2005; Gertner and VanLehn 2000; Koedinger et al. 1997; Lesgold, Lajoie, Bunzo and Eggan 1992; Sleeman and Brown 1982; Woolf 2009]. It was around this time that *Affective Computing* was beginning to emerge as a new and exciting research area. Affective Computing focuses on creating technologies that can monitor and appropriately respond to the affective states of the user [Picard 1997; Picard 2010]. Affective computing is a subfield of human-computer interaction (HCI), where the affective states of a user (feelings, moods, emotions) are incorporated into the decision cycle of the interface in an attempt to develop more effective, user-friendly, and naturalistic applications.

Affective computing is particularly relevant to ITSs because the ITSs that were developed prior to 2000 primarily focused on students' cognitive states and sometimes motivational levels [del Soldato and du Boulay 1995; du Boulay 2011]. This is a critical limitation because learning and problem solving are rife with emotional experiences [Calvo and D'Mello 2011; D'Mello and Graesser 2012; Meyer and Turner 2006; Pekrun et al. 2010; Pekrun and Stephens 2012]. The inextricable link between affect and learning suggests that ITSs can be more than mere cognitive machines, and they should be affective processors as well. Affect-sensitivity is important for ITSs that aspire to model human tutors because it has been claimed that expert teachers are able to recognize a student's emotional state and respond in an appropriate manner that has a positive impact on the learning process [Goleman 1995; Lepper and Woolverton 2002]. An affect-sensitive ITS would incorporate assessments of students' cognitive and affective states into its pedagogical and motivational strategies in order to keep students engaged, boost self-confidence, heighten interest, and presumably maximize learning [Calvo and D'Mello 2011].

A number of research groups have recently focused on building learning environments that detect and respond to affective states such as boredom, confusion, frustration, and anxiety [Afzal and Robinson 2011; Burleson and Picard 2007; Chaffar et al. 2009; Conati and Maclaren 2009; D'Mello and Graesser 2010; D'Mello, Lehman, Sullins, Daigle, Combs, Vogt, Perkins and Graesser 2010d; Forbes-Riley et al. 2008; Robison et al. 2009; Woolf et al. 2010]. These systems use state-of-the art sensing technologies and machine learning techniques to automatically detect student affect by monitoring facial-features, speech contours, body language, interaction logs, language, and peripheral physiology (e.g., electromyography, galvanic skin response) (see [Calvo and D'Mello 2010] for an overview). These affect-sensitive systems then alter their pedagogical and motivational strategies in a manner that is dynamically responsive to the sensed affective states. Some of the implemented responses to student affect include affect mirroring [Burleson and Picard 2007], empathetic responses [Woolf, Arroyo, Muldner, Burleson, Cooper, Dolan and Christopherson 2010], and a combination of empathy, encouragement, and incremental challenge [D'Mello, Lehman, Sullins, Daigle, Combs, Vogt, Perkins and Graesser 2010d]. Section 5 describes one such system, namely the affect-sensitive AutoTutor or Affective AutoTutor.

3. AUTOTUTOR

AutoTutor simulates a human tutor by holding a conversation with students in natural language. Students type in their contributions through a keyboard in most applications. However, we have developed a version that handles spoken input from the student through the Dragon Naturally Speaking ™ (version 6) speech recognition system (See Section 7.1). AutoTutor communicates through an animated conversational agent with speech, facial expressions, and some rudimentary gestures (see Section 0 below).

AutoTutor has been implemented and tested for the domains of Newtonian physics, computer literacy (the Internet, operating systems, and hardware), and scientific methods (see Section 0 for alternate versions of AutoTutor). AutoTutor's tutorial sessions are typically geared to promote conceptual thinking and deep reasoning rather than memorization of definitions and facts. However, there is nothing in its design to prevent AutoTutor from helping students with the acquisition of domain-specific facts. In essence, it is possible to create tutorial problems that differentially map onto taxonomies of cognitive difficulty [Bloom 1956; Wakefield 1996]. For example, there are the three levels of difficulty (easy, medium, difficult) for AutoTutor's computer literacy problems (see [Person et al. 2003] for details).

As with most ITSs, AutoTutor's behavior can be characterized by an outer and an inner loop [VanLehn 2006]. The outer loop of AutoTutor consists of a series of didactic lessons and challenging problems or *main questions* (such as *why, how, what-if*). An example main question is "When you turn on the computer, how is the operating system first activated and loaded into RAM?" The order of lessons, problems, and questions can be dynamically selected based on the profile of student abilities, but the order is fixed in most versions of AutoTutor we have developed. The interactive dialogue occurs during the construction of an answer to a main question but not during the didactic delivery of information (e.g., reading text, viewing a diagram). The answer to a question (or solution to a problem) requires several sentences of information in an ideal answer. AutoTutor assists the student in constructing the answer after the student enters an initial response.

The inner loop of AutoTutor consists of this collaborative interaction while answering a question (or solving a problem). It is this inner loop that is the distinctive hallmark of AutoTutor. The tutor draws out more of the student's knowledge (through hints and prompts), helps fill in missing information, repairs misconceptions, and answers student questions. The inner loop dialogue between AutoTutor and the student takes approximately 100 dialogue turns to answer a single challenging question, approximately the length of a conversation with a human tutor [Graesser, Person and Magliano 1995].

3.1 What does AutoTutor do?

Perhaps the best way to convey what AutoTutor does in the inner loop is through an example conversational dialogue. The excerpts below were extracted from an actual conversation with a student on the topic of conceptual physics. The conversation begins with a main question selected by AutoTutor.

> **AutoTutor***: When a car without headrests on the seats is struck from behind, the passengers often suffer neck injuries. Explain why.* (**main question**)
>
> **Student***: I don't know.* (**metacognitive speech act**: This expression by the student is classified as a metacognitive speech act because it conveys what the student knows about his/her own knowledge.)
>
> **AutoTutor***: Could you at least try to give me an answer? That way, we'll have something as a starting place to work from.* (**pump**: This indirect dialogue move encourages the student to say more, but does not direct the student to express particular information.)

As the dialogue continues, AutoTutor detects that an important expected piece of information is missing (i.e., the impact of the collision on the neck).

> **AutoTutor*:*** *Let's try this together.* (**discourse marker**: This is a discourse marker that introduces or connects dialogue moves within a turn.)
>
> *When the collision provides a large force to accelerate the head, what could happen to the neck*? (**hint**: This dialogue move leads the student to articulate a particular sentence, clause, or proposition.)

At one point in the dialogue, the student asks a question, so AutoTutor needs to provide an answer. AutoTutor can answer some classes of questions (such as definitional questions), but not all questions, so AutoTutor needs to productively manage the dialogue surrounding student question asking.

> **Student***: What is a force?*
>
> **AutoTutor**: *I'll take a crack at answering your question.* (**discourse marker**)
>
> *Force is any influence that tends to accelerate an object; a push or pull; measured in newtons. Force is a vector quantity.* (**answer**: AutoTutor answers the student's direct question. Answers are either accessed from electronic glossaries or a paragraph is dynamically selected from a textbook.)
>
> *Was that what you were looking for?* (**question solicitation**: AutoTutor allows the student the opportunity to rephrase his question.)

As the dialogue continues AutoTutor believes it has detected a misconception.

> **AutoTutor**: *See if you know the answer to this.* (**discourse marker**: AutoTutor prepares the student for a direct question.)
>
> *When the head and body are moving at the same acceleration, they are moving at the same ________?* (**prompt**: This more direct dialogue move encourages the student to fill in a particular word or phrase of information, in this case the word "rate." Prompts are frequently needed for students who have low verbal fluency in order to elicit a particular noun, verb, or adjective that is missing in an answer.)
>
> **Student**: *distance*
>
> **AutoTutor:** *Moving at the same rate.* (**prompt completion**: AutoTutor splices in the correct information if the student provides an incorrect answer to its prompt. The misconception is eventually corrected with an assertion as the dialogue advances.)

At this point, the dialogue continues until all of the relevant expected answers are covered in the conversation. The conversation then ends with a summary answer by the tutor, or the tutor might prompt the student to construct a summary.

The example illustrates the mixed-initiative dialogue of AutoTutor. AutoTutor attempts to interpret or otherwise handle any question, assertion, comment, or extraneous speech act that the student expresses. Each turn of AutoTutor requires the generation of one or more dialogue moves that adaptively respond to what the student expresses, that advance the conversation in a constructive manner, that cover good answer information, or that correct misconceptions. The tutor's dialogue moves within a turn are connected by discourse markers, as illustrated in the example. Some dialogue moves are responsive to the student's preceding turn, such as short feedback (positive, neutral, versus negative), answers to student questions, and corrections of student misconceptions. Other dialogue moves push the dialogue forward in an attempt to cover the expected answers to the main question.

3.2 Dialogue Structure

The dialogue structure of AutoTutor is similar to the dialogue patterns of human tutors. Graesser and Person analyzed dialogue patterns of typical human tutors in middle school and in college [Graesser and Person 1994; Graesser, Person and Magliano 1995]. Similar analyses have been conducted by other researchers on naturalistic tutoring corpora [Chi et al. 2004; D'Mello, Olney and Person 2010e; Evens and Michael 2006; Litman et al. 2006; VanLehn et al. 2003]. The following dialogue structures are prominent in human tutors and are implemented in AutoTutor: (a) a curriculum script with didactic content and problems (i.e., difficult tasks or questions), (b) a 5-step Tutoring Frame, (c) Expectation and Misconception Tailored (EMT) dialogue, and (d) Conversational Turn Management.

Curriculum script. The tutor covers a curriculum with didactic content and a set of questions or problems that address the content. Didactic content can be presented in a mini-lecture, hopefully at the appropriate time for each individual student. The questions/problems require the student to actively apply their knowledge. The curriculum script includes expected answers, misconceptions, hints, prompt questions, and other inner loop information.

5-Step tutoring frame. When a challenging main question (or problem) is selected to work on, the question is answered through an interaction that is structured by a 5-Step Tutoring Frame. The 5 steps are: (1) the tutor presents a main question, (2) the student gives an initial answer, (3) the tutor gives short feedback on the quality of the student's initial answer, (4) the tutor and student collaboratively improve on the answer in a turn-by-turn dialogue that may be lengthy, and (5) the tutor evaluates whether the student understands (e.g., asking "Do you understand?" or testing with a follow-up task). In the spirit of constructivism, this 5-step tutoring frame involves collaborative discussion, joint action, and encouragement for the student to construct knowledge rather than merely receiving knowledge.

Expectation and misconception tailored (EMT) dialogue. Human tutors typically have a list of *expectations* (i.e. anticipated good answers or steps in a procedure) and a list of anticipated *misconceptions* (incorrect information) associated with each main question. They want the expectation content covered in order to handle the main question that is selected. The tutor guides the student in articulating the expectations through a number of dialogue moves, namely *pumps* ("What else?"), *hints* ("What about X?"), *prompt questions* to extract specific information from students ("X is a type of what?"), *assertions* that capture particular expectations ("X is a type of Y"), and *answers* to students' questions.

As the dialogue progresses, tutors tend to lead more while trying to get the student to articulate an expectation. They start with a pump and then move to a hint if the pump fails, followed by a prompt question and an assertion if students fail to articulate the expectation. The pump → hint → prompt → assertion cycle is implemented by AutoTutor to encourage the student to articulate the answer and cover expectations. The correct answers are eventually covered and the misconceptions are hopefully corrected.

Conversational turn management. Human tutors structure their conversational turns systematically. Nearly every turn of the tutor has three information slots. The first slot of most turns is feedback on the quality of the student's last turn. This feedback is either positive (e.g., "very good", "yeah"), neutral (e.g., "uh huh", "I see"), or negative (e.g., "not quite", "not really"). The second slot advances the interaction with a prompt for specific information, a hint, an assertion with correct information, a correction of misconceptions, or an answer to the student's question. The third slot is a cue for the floor to shift from the tutor as the speaker to the student. For example, AutoTutor ends each turn with a question or a gesture to cue the student to do the talking. Otherwise the student and AutoTutor are at a standstill waiting for the other to take the next turn.

3.3 Monitoring Students' Cognitive States

One of the central questions is how well the tutor can track the psychological states of the student as the tutor implements tutoring strategies. Available evidence suggests that novice human tutors are not able to conduct student modeling at a fine-grained level [Chi, Siler and Jeong 2004; Graesser et al. 2009; Person et al. 1994]. They are limited to performing approximate assessments rather than fine-grain assessments. On the other hand, there is evidence to suggest that some of the more accomplished (or expert) human tutors perform fine-grained student modeling because they are quite adept at assessing student knowledge [D'Mello et al. 2010c; Lepper and Woolverton 2002]. Computers can potentially show advantages over novice human tutors to the extent that artificial intelligence algorithms can be used to accurately conduct student modeling and generate intelligent responses.

Semantic matching algorithms. Student modeling in the inner loop is executed after each turn and consists of comparing what the student expresses in language with the list of expectations and misconceptions associated with a main question. This requires *semantic matching* algorithms that compare the student input with AutoTutor's expectations and misconceptions. It is widely accepted that natural language is imprecise, fragmentary, vague, ungrammatical, and elliptical, so it would not be prudent to rely entirely on semantically well-formed semantic matches. AutoTutor therefore incorporates several semantic evaluation algorithms when performing these matches, but notably Latent Semantic Analysis [Landauer et al. 2007], regular expressions, content word overlap metrics (that have higher weight for low frequency words than high frequency words) [D'Mello et al. 2010b], and occasionally logical entailment [Rus and Graesser 2007]. LSA is a statistical technique that measures the conceptual similarity of two text sources. In this similarity matching algorithm, a vector representing the semantic content of the contribution is created and compared to vectors that represent the semantic content of expectations and misconceptions. The cosine between the two vectors is calculated to produce a match similarity score from 0 to 1 (negative cosines are rare and are converted to 0 in AutoTutor).

Covering an expectation and detecting a misconception. Early versions of AutoTutor relied exclusively on LSA in their semantic evaluation of student input. The LSA algorithm in AutoTutor computed the extent to which the information within the student turns (i.e., an individual turn, a combination of turns, or collective sequence of turns) semantically matches each expectation in the ideal answer.

Expectation E_i is considered covered if the content of the student's cumulative set of turns meets or exceeds a threshold T in its LSA cosine value (which varies from near 0 to 1). That is, E_i is covered if the cosine match between E_i and the student input I (including turns 1 though N) is high enough: cosine $(E_i, I) \geq T$. The threshold has varied between .40 and .75 in previous instantiations of AutoTutor.

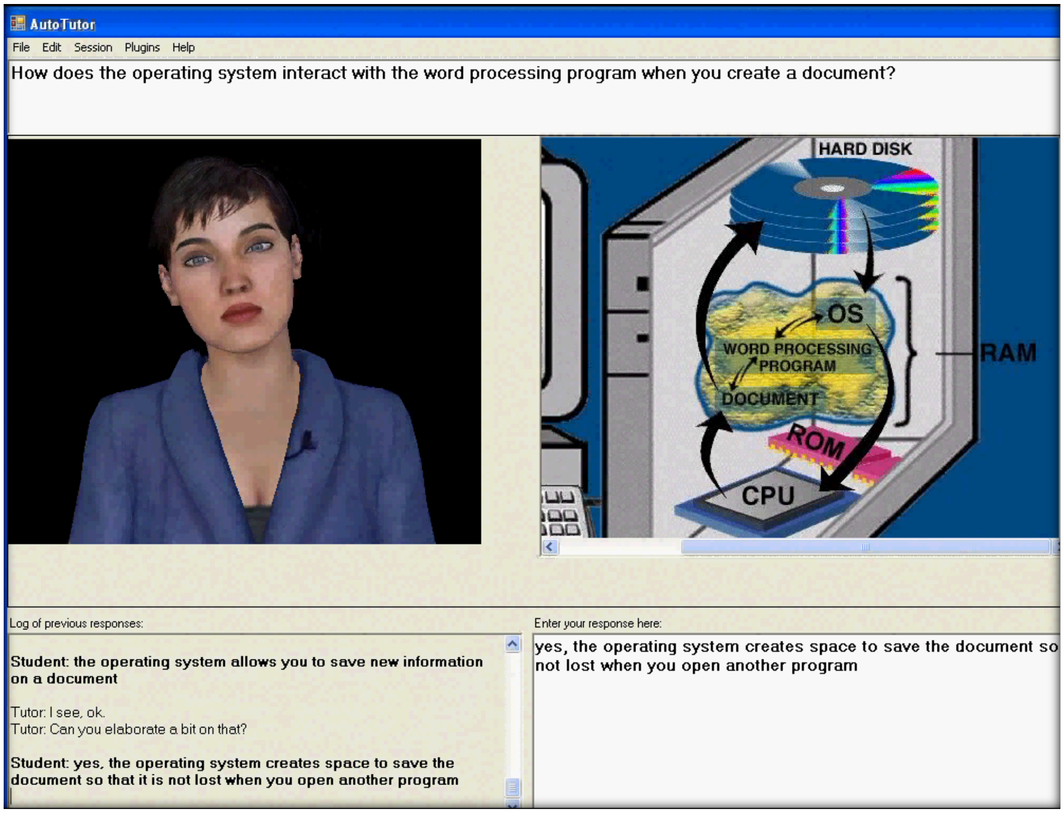
Each expectation E_i has an associated family of prompts and hints to get the student to fill in most or all of the content words and propositions in E_i. Prompts and hints are selected to maximize an increase in the LSA cosine match score (hereafter called the *match score*) when answered successfully. Stated differently, hints and prompts are selected to maximize pattern completion. Sometimes the student expresses misconceptions during the dialogue. This happens when the student input I matches a misconception M with a sufficiently high match score. At that point AutoTutor corrects the misconception and goes on.

Selecting the next expectation to cover. During the course of the dialogue and student modeling, the system periodically identifies a missing expectation and posts the goal of covering the expectation. When expectation E_i is missed (and therefore posted), AutoTutor attempts to get the student to articulate it by generating hints and prompts affiliated with E_i to help the student fill in missing words and propositions. The selection of the next E_i to cover follows the principle of the *zone of proximal development* or what some call frontier learning [Brown et al. 1998; Vygotsky 1986]. That is, AutoTutor builds on what the student has managed to articulate. More formally, AutoTutor selects the next E_i from the set of expectations that (a) has the highest match score and (b) has a subthreshold match score (i. e., the expectation has not yet been covered). This *subthreshold expectation selection* algorithm assumes that the expectations should not be covered in a prescribed sequential order. However, ordering constraints may also be considered in a *sequential expectation selection* algorithm. Some subject matters have ordering constraints but others do not.

Student modeling across turns. The above specification for a single student-tutor turn pair does not tell the whole story, however. If the student model was completely rebuilt on each student turn, the semantic matches would wildly vary, representing the vicissitudes of the student's ability to provide content in response to different tutor dialogue moves. Additionally, the student's responses to AutoTutor's pumps and hints are typically only a few words in length. This relatively low verbosity introduces problems with the fidelity of the semantic matching algorithms, especially LSA. Therefore, the semantic match is calculated using both the current student response alone (local assessment) and the current response concatenated with all previous student responses for the current problem (global assessment). This global assessment represents AutoTutor's model of student knowledge across turns for a particular main problem. It moves beyond short student responses and helps AutoTutor maintain continuity of the student model between turns.

3.4 AutoTutor Interface

AutoTutor's interface varies across versions, but most versions have the five major windows shown in Figure 1. Window 1 (top of screen) is the main question that stays on the computer screen throughout the conversation about the question. Window 2 (left middle) is the animated conversational agent that speaks the content of AutoTutor's turns. Window 3 (right middle) is either blank or has auxiliary diagrams. Window 4 (right bottom) displays the students' answers as they type them in. Window 5 (left bottom) displays the dialogue history of the student and the tutor.


AutoTutor
File Edit Session Plugins Help
HARD DISK
OS
WORD PROCESSING PROGRAM
DOCUMENT
RAM
ROM
CPU
Log of previous responses:
Enter your response here:

How does the operating system interact with the word processing program when you create a document?

Student: the operating system allows you to save new information on a document

Tutor: I see, ok.
Tutor: Can you elaborate a bit on that?

Student: yes, the operating system creates space to save the document so that it is not lost when you open another program

yes, the operating system creates space to save the document so not lost when you open another program

Figure 1. AutoTutor interface

3.5 AutoTutor Architecture

The major components of AutoTutor's architecture are presented in Figure 2. AutoTutor operates as a distributed client-server application implementing a transient asynchronous mode of communication. The client and the server are written in the C# language within the Microsoft .NET framework. The subsequent discussion provides a very brief overview of the architecture. Additional details can be found in Graesser et al. [2005].

Speech act classification. The bottom left of the figure depicts the student entering information via the user interface. The information in each student turn is segmented into speech acts, based on punctuation and (in some systems) a syntactic parser. Each speech act is assigned to one of approximately 20 speech act categories. These categories include assertions, 16 different categories of questions, short responses (e.g., “yeah”, “right”), metacognitive expressions (e.g., “I don't understand”, “I see”), and metacommunicative expressions (e.g., “What did you say?”). The accuracy of classifying the student speech acts into categories varies from 0.87 - 0.96 percent [Olney et al. 2003].

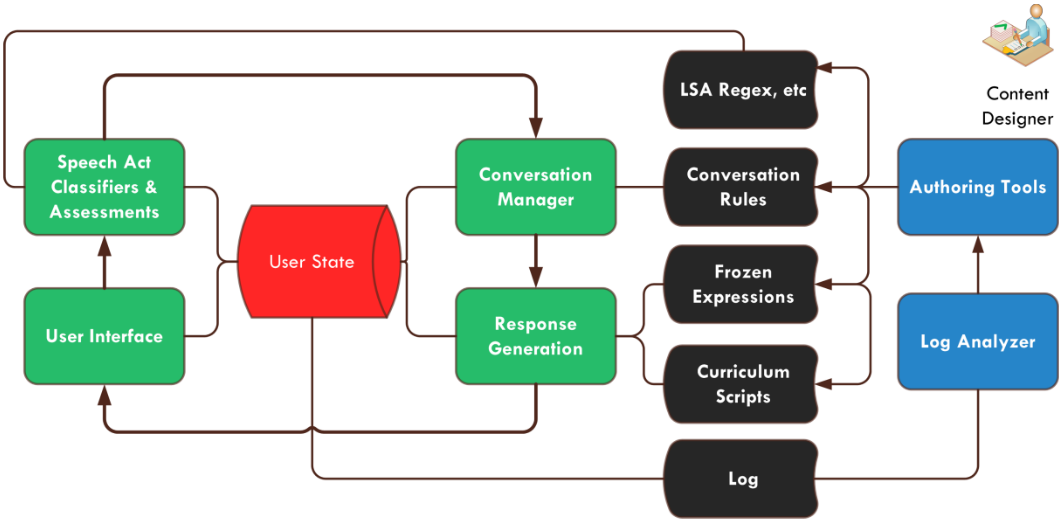


Figure 2. AutoTutor's architecture

Conversational management and response generation. The speech acts expressed by the student on any given turn *N* constrain AutoTutor's conversation management of turn *N+1*. If the student asks a question, AutoTutor needs to answer it if it has an answer, or otherwise it (a) generates dialogue moves to put the onus on the student to find an answer (e.g., "Good question. How would you answer it?") or (b) generates dialogue moves that evade getting an answer (e.g., "Good question, but I cannot answer it now. Let's move on.[2]"). If the student provides a metacognitive response (e.g., "I'm lost, I don't know"), then AutoTutor acknowledges this and presents a hint to advance the dialogue in productive avenues. Most of the student responses are answers to the tutor's hints, prompts, and pumps. These are evaluated on quality (see Section 3.3) and this evaluation drives the pump → hint → prompt → assertion cycles.

The *Conversation Manager* is sensitive to the student's speech acts and is generally responsible for keeping track of the tutorial dialogue. The Conversation Manager can be represented by a set of "*if* <state> *then* <action>" production rules [Anderson and Gluck 2001] or of a finite state transition network [Graesser et al. 2001a]. The Conversation Manager passes information to the *Response Generator*, which generates the actual text of the dialogue moves and adds appropriate discourse markers. This content is expressed either via text or by an animated conversational agent that is displayed on the interface (see Section 0).

Data structures and databases. As depicted in the figure, AutoTutor has a repository of different static data structures that can be created and updated with authoring tools. First, most versions of AutoTutor represent world knowledge as LSA spaces, but some versions of AutoTutor or its progeny have incorporated other forms of world knowledge representation, such as textbooks, glossaries, and conceptual graph structures. Second, there are *Conversation Rules* that are represented as production rules, finite-state transition networks, or recursive augmented state transition networks. Third, there are different categories of *Frozen Expressions* that have different discourse functions. For example, there are different ways for AutoTutor to express positive feedback (e.g., "yes", "yeah", "good", "great", "fantastic",

---

[2] AutoTutor does not revisit the student's question, although the question sometimes naturally gets answered over the course of the dialogue.

"right on") and different ways that the student can express metacommunicative speech acts (e.g., "What did you say?" "Please repeat. I did not hear that").

The *Curriculum Script*, as briefly described in Section 3.2, is a critical component of the AutoTutor architecture. Each script contains the content associated with a question or problem. For each, there is (1) the ideal answer, (2) a set of expectations, (3) families of potential hints, correct hint responses, prompts, correct prompt responses, and assertions associated with each expectation, (4) a set of misconceptions and corrections for each misconception, (5) a set of key words and functional synonyms, (6) a summary, and (7) markup language for the speech generator and gesture generator for components in (1) through (6) that require actions by the animated agents. AutoTutor is equipped with a *Script Authoring Tool* [Susarla et al. 2003] to enable subject-matter experts to easily create the content of the curriculum script. The complete *Curriculum Script* for one computer literacy question in published in the Appendix of [Person, Graesser, Kreuz and Pomeroy 2003].

All of the information collected during the AutoTutor-student interaction is stored in *Log* files. These files are fed into the *Log Analyzer* that can be inspected by the researcher and can inform the lesson planner or knowledge engineer who uses the *Authoring Tools* to create new content or edit existing content. These modules are, of course, standard for all learning management systems.

4. EVALUATIONS OF AUTOTUTOR

AutoTutor has been evaluated along a number of dimensions such as the quality of its dialogue, the accuracy by which it evaluates student responses, its student model, and most importantly its efficacy in promoting learning gains. This section provides a synthesis of the major finding along each of these four dimensions.

4.1 Evaluating AutoTutor's Conversational Smoothness

We performed a *bystander Turing test* to evaluate the naturalness of AutoTutor's dialogue moves [Person and Graesser 2002]. We randomly selected 144 tutor moves in the tutorial dialogs between students and AutoTutor. We asked six human tutors to fill in what they would say at these 144 points. So, at each of these 144 tutor turns, we had what the human tutor generated and what AutoTutor generated. We subsequently tested whether a group of students could discriminate between dialogue moves that were generated by a human versus a computer; half in fact were generated by the human tutors and half were by AutoTutor. We found that the bystander students were unable to discriminate whether particular dialogue moves had been generated by a computer versus a human; the *d'* discrimination scores were actually a bit negative (-.08), but not significantly. This rather impressive outcome supports the claim that AutoTutor is a good simulation of human tutorial dialogue.

4.2 Evaluating AutoTutor's Semantic Matching Algorithms

We have analyzed the accuracy of the match evaluation scores by comparing AutoTutor's scores to judgments of subject matter experts [Graesser et al. 2007b; Graesser et al. 2000]. For example, we have analyzed the complete answers that students gave as an answer to one of the challenging physics questions, recorded AutoTutor's match evaluation score for each expectation/misconception, and collected ratings from 5 expert physicists as to whether each expectation/misconception was present in the student answers. The correlations between these match evaluation scores and expert ratings have varied between 0.35 and 0.50, depending on the criterion, semantic algorithm, and other details that need not be considered here.

One expert physicist rated the degree to which particular speech acts expressed during AutoTutor training matched particular expectations. These judgments were made on a sample of 25 physics expectations and 5 randomly sampled student answers per expectation, yielding a total of 125 pairs of expressions. The question is how well the expert ratings correlate with match evaluation score for the relevant expectation. We found that the correlation between an expert judge's rating and the match evaluation score was modest ($r = .29$), but significant in accounting for the 125 items.

4.3 Evaluating AutoTutor's Student Modeling

The accuracy of the student model algorithms have been evaluated over the years. In one analysis of conceptual physics, we collected pretest scores on a psychometrically validated test called the Force Concept Inventory [Hestenes et al. 1992]. If AutoTutor is performing effective user modeling, then the dialogue moves selected by AutoTutor should be correlated with the students' prior knowledge of physics. Such predictions held up when we analyzed the dialogue moves of AutoTutor as a function of students' prior knowledge [Jackson and Graesser 2006]. For example, the short feedback that AutoTutor provides after the students' turns is either positive, neutral, or negative. The students' physics knowledge had a significant positive correlation with positive feedback moves ($r = .38$) and a negative correlation with negative feedback ($r = -.37$). Another example applies to the corrections that AutoTutor made when identifying student errors and misconceptions. The correlation between prior knowledge and corrections was negative ($r = -.24$), and marginally significant.

Yet another example pertains to the four dialogue move categories that attempt to cover the content of the expectations in the curriculum script: Pumps, hints, prompts, and assertions. The proportion of dialogue moves in these categories should be sensitive to the student's knowledge of physics. There is a continuum from the student-provided information to tutor-provided information as we move from pumps, to hints, to prompts, to assertions. The correlations with student knowledge reflected this continuum perfectly, with values of 0.49, 0.24, -0.19, and -0.40, respectively. Thus, for students with more knowledge of physics, all AutoTutor needs to do is primarily pump and hint, thereby encouraging or nudging the student to supply the answer to the question and articulate the expectations. For students with less knowledge of physics, AutoTutor needs to generate prompts for specific words or to assert the correct information, thereby extracting knowledge piecemeal or telling the student the correct information. These results support the claim that AutoTutor performs user modeling with some degree of accuracy and adaptively responds to students' knowledge levels.

4.4 Learning Gains Produced by AutoTutor

Perhaps the most important question is whether AutoTutor helps students learn. The learning gains of AutoTutor have been evaluated in over 20 experiments since its inception. Training times in these studies varied from 30 minutes to 4 hours and tutorial sessions were sometimes split across multiple days. Assessments of AutoTutor on learning gains have shown effect sizes of approximately 0.8 sigma in the areas of computer literacy [Graesser, Lu, Jackson, Mitchell, Ventura, Olney and Louwerse 2004] and Newtonian physics [VanLehn, Graesser, Jackson, Jordan, Olney and Rose 2007]. These effect sizes place AutoTutor somewhere between an untrained human tutor [Cohen, Kulik and Kulik 1982] and an ITS with ideal tutoring strategies [Corbett 2001]. Some of the hypothesized ideal tutoring strategies include Socratic teaching, delayed feedback, motivational support [Lepper and Chabay 1988; Lepper and Woolverton 2002]. AutoTutor improves learning between 0 and 2.1 sigma (a mean of 0.8), depending on the learning performance measure, the comparison condition, the subject matter, and the version of AutoTutor.

Measures of learning have varied in scope, depth, difficulty, and open-endedness. They have included: (1) multiple choice questions on shallow knowledge that tap definitions, facts and properties of concepts, (2) multiple choice questions on deep knowledge that tap causal reasoning, justifications of claims, and functional underpinnings of procedures, (3) essay quality when students attempt to answer challenging problems, (4) a cloze task that has students fill in missing words of texts that articulate explanatory reasoning on the subject matter, and (5) performance on problems that require problem-solving.

Assessments of learning gains obviously depend on the comparison conditions. The learning gains are approximately 0.8 for AutoTutor compared to a do-nothing control or a condition of reading from a textbook on the same topics for an equivalent amount of time. The learning gains are approximately the same for AutoTutor and an expert human tutor who interacts with the student by computer-mediated communication (as opposed to face-to-face).

The largest learning gains from AutoTutor have been on deep-reasoning measures rather than measures of shallow knowledge [VanLehn, Graesser, Jackson, Jordan, Olney and Rose 2007]. AutoTutor is most effective when there is an intermediate gap between the student's prior knowledge and the ideal answers of AutoTutor; AutoTutor is not particularly effective in facilitating learning in students with high domain knowledge, nor when the material is too much over the student's head.

4.5 Limitations and Potential Areas for Improvement

The assessments point to the successes of AutoTutor, but it is important also to acknowledge some of its limitations. One limitation is that the conversational dialogue may have minimal incremental gains on learning when the exchange is time-consuming and the knowledge covered is shallow rather than deep. The conversational interaction is tedious for some students and even irritating for a small percentage. A second limitation is that students lose patience with AutoTutor when the conversation breaks down. Such breakdowns occur when the student modeling is imperfect, the curriculum script is incomplete, student speech acts are misclassified, and AutoTutor is viewed as being unresponsive to what the student is saying. A third limitation is that AutoTutor can correctly answer only a modest proportion of student questions so students eventually stop asking them[3].

One important future direction is to improve the student modeling and conversational facilities of AutoTutor in order to minimize some of its persistent blemishes. This can be accomplished in a number of ways. For example, there can be checks in the authoring tools to make sure that the content is complete when it is prepared by the author of the curriculum scripts. Another direction is to develop more sophisticated question answering facilities so that AutoTutor can answer a broad diversity of questions. This would contribute to mixed-initiative dialogue and put more control in the hands of the student.

5. AFFECTIVE AUTOTUTOR

AutoTutor is quite effective in helping students learn by modeling and responding to their cognitive states. However, this is only one part of the story because learning consists of a complex interplay between cognition, emotions, and learning [Snow et al. 1996]. We have recently developed two new versions of AutoTutor that detect and respond to students' affective and cognitive states [D'Mello et al. 2008b; D'Mello et al. 2009]. These affect-sensitive versions of AutoTutor, called the Supportive and

---

[3] Some of these communication failures were systematically tracked in a study that compared the traditional typed input version of AutoTutor with a spoken-input version of the tutor. Please see [D'Mello, King and Graesser 2010b] for details.

Shakeup tutors, are collectively referred to as Affective AutoTutor. They have a set of production rules that were designed to map dynamic assessments of the student's cognitive and affective states with tutor actions to address the presence of boredom, confusion, and frustration.

The achievement of an affect-sensitive tutorial interaction engages the tutor and student in an *affective loop* [Conati et al. 2005]. This loop includes the real-time *detection* of the affective states that are relevant to learning, the *selection* of appropriate tutor actions that maximize learning while influencing the student's affect, and the *synthesis* of emotional expressions by the tutor as it attempts to engage the student in a more human-like, naturalistic manner.

The affective loop in an integrated system can be viewed from the perspective of either the student or the tutor. The student-centric view consists of analyzing the prominent affective states in the student, assessing their potential impact on learning, identifying how these states are manifested in the student, and developing an automatic affect detection system. The tutor-centric view explores how good human tutors or theoretical ideal tutors adapt their instructional agenda to encompass the emotions of the student. This expert knowledge is then transferred to computer tutors such as Affective AutoTutor. Embodied conversational agents that simulate human tutors are programmed to synthesize affective elements through the generation of facial expressions, the inflection of speech, and the modulation of posture.

5.1 Identifying Affective States

There is the important issue of identifying the affective states that students experience during interactions with AutoTutor and other learning environments. One possibility is that the *basic emotions* (anger, sadness, fear, disgust, happiness, and surprise) [Ekman 1992] constitute students' primary emotional reactions. Alternatively, the so called *academic emotions* or *learning-centered* emotions (e.g., anxiety, boredom) might be relevant in learning contexts [Pekrun and Stephens 2012; Schutz and Pekrun 2007]. We addressed this fundamental question by conducting a number of studies that aimed at identifying the affective states that students typically experience while interacting with AutoTutor, with the expectation that these findings will generalize to other learning environments [Baker et al. 2010; D'Mello in review].

In the *Observational* study, five trained judges observed the affective states (boredom, confusion, frustration, eureka, flow/engagement, versus neutral) of 34 students who were learning introductory computer literacy with AutoTutor [Craig et al. 2004]. In the *Emote-Aloud* study, seven college students verbalized their affective states while interacting with AutoTutor [D'Mello et al. 2006]. The *Multiple-Judge* study consisted of 28 students completing a 32-minute session with AutoTutor, after which their affective states were judged by the students themselves, untrained peers, and two trained judges. Judgments were based on videos of students' faces and computer screens which were recorded during the tutorial session [Graesser et al. 2006]. The *Speech Recognition* study was similar to the multiple-judge study with the exception that students spoke their responses to the AutoTutor system instead of typing them (see Section 7.1). Retrospective self-reports by the students constituted the primary affect measure in this study [Graesser et al. 2007a]. The *Physiological* study also implemented the retrospective affect judgment procedure, but the students were 27 engineering students from an Australian University [Pour et al. 2010] instead of the undergraduate psychology students from the U.S. who comprised the samples in the previous four studies.

When averaged across studies, flow/engagement was the most frequent state, comprising 24% of the observations. Boredom and confusion were the second most frequent states (18% and 17%, respectively) followed by frustration (13%). Neutral was reported for 19% of the observations, while delight (6%) and surprise (3%) were rare. Indeed, boredom, flow/engagement, confusion, and frustration are more frequent affective states that students experience during interactions with AutoTutor.

Although the present set of studies did not directly compare the occurrence of these learning-centered affective states with the basic emotions, we are currently in the process of conducting a meta-analysis of 21 studies that have tracked the incidence of both the learning-centered (e.g., boredom, confusion, flow/engagement, frustration) and basic emotions [D'Mello in review]. Preliminary results of the meta-analysis indicate that with the exception of happiness, which occurs with some frequency, the remaining basic emotions were considerably rare when compared to the learning-centered affective states. The basic emotions have claimed center-stage of most emotion research in the last four decades, but these results suggest that they might not be relevant to learning, at least for the short learning sessions (30 minutes – 1.5 hours) of the studies that were analyzed. In contrast, confusion, frustration, flow/engagement, and boredom were the prevalent emotions, indicating that it is critically important for Affective AutoTutor to respond to these states.

5.2 Detecting Affective States

Our affect detection system monitors conversational cues [D'Mello, Craig, Sullins and Graesser 2006; D'Mello et al. 2008a], gross body language [D'Mello and Graesser 2009; D’Mello et al. 2007], and facial features [D'Mello et al. 2007; McDaniel et al. 2007] (see Figure 3). The classifier was trained on data obtained in a study that involved synchronization and data recording of the sensors while 28 students interacted with AutoTutor [D'Mello and Graesser 2010]. Manually annotated affect labels, which are required for the supervised learning systems, were obtained by multiple human judges including the student (self judgments), an untrained peer, and two trained judges (see Multiple-Judge study in Section 5.1 and [D'Mello and Graesser 2010] for additional methodological details).

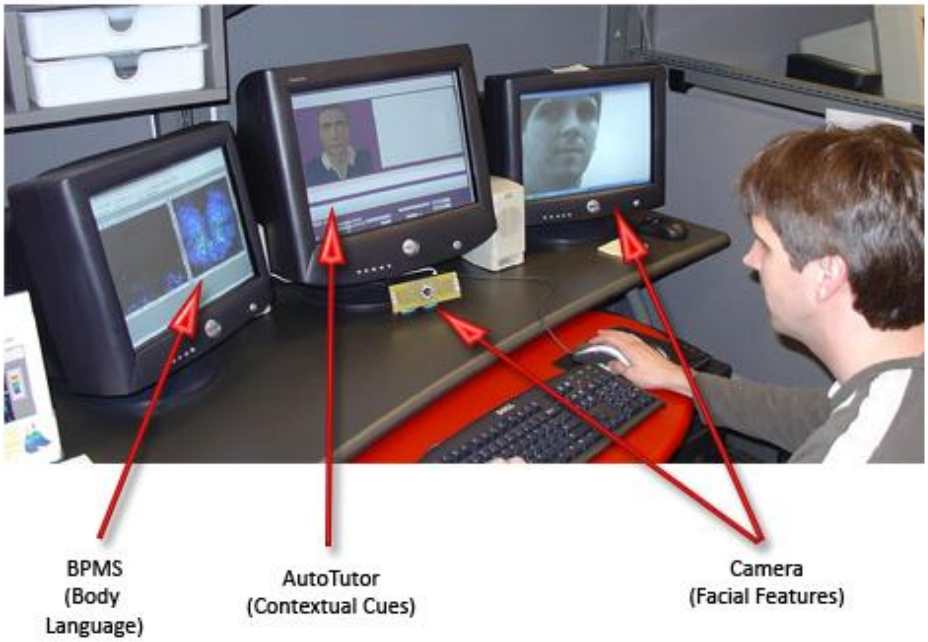


Figure 3. Automated affect sensing

Conversational cues (dialogue features). A one-on-one tutoring session with AutoTutor yields a rich trace of contextual information, characteristics of the student, episodes during the coverage of the topic, and social dynamics between the tutor and student. These conversational cues cover a broad and deep feature set that includes assessments of deep meaning, world knowledge, and pragmatic aspects of communication. Therefore, several conversational features and discourse markers (collectively called dialogue features) were extracted from AutoTutor's log files and were utilized to infer the student's affect. The dialogue features were computed for each student-tutor turn (i.e. student submits response, tutor provides feedback, tutor presents next question). They included *temporal* features (e.g. time on problem, response time), assessments of *response verbosity* (e.g. number of characters, speech act), assessments of the *conceptual quality* of the student's response obtained by Latent Semantic Analysis (LSA), *conversational directness* (i.e. how much information the tutor is explicitly providing to the student), and *tutor feedback* (negative, neutral, positive). The full list of features is can be found in D'Mello et al., [2008].

Gross body language (posture features). The Body Posture Measurement System (BPMS), developed by Tekscan™, was used to monitor the gross body language of students during tutorial sessions with AutoTutor (see left monitor of Figure 3). The BPMS consists of a thin-film pressure pad (or mat) that can be mounted on a variety of surfaces. The output of the BPMS system consisted of two 38 ×41 matrices (for the back and seat) with each cell in the matrix corresponding to the amount of pressure exerted on the corresponding element in the sensor grid.

We relied on an *attentive-arousal* framework [Bull 1987] to interpret relationships between the posture features and the affective states of the student. One can think of heightened pressure in the seat as resonating with a tendency to position one's body towards the source of stimulation (i.e., high attentiveness since the student is positioning his or her body towards the AutoTutor interface, or a short distance between the nose and the screen). On the other hand, an increase in pressure on the back of the chair suggests that the student is leaning back and detaching one's self from the stimulus (low attentiveness). Arousal was operationally defined by the rate of change of pressure exerted on the back and the seat of the pressure sensitive chair, and is similar to the degree of activation. In addition to these primary features, we also tracked changes in body position and arousal before and after an emotional episode (see [D'Mello and Graesser 2009] for details).

Facial feature tracking. We use the Mindreader system [el Kaliouby and Robinson 2005a; el Kaliouby and Robinson 2005b] for fully automated facial feature tracking. The Mindreader system uses a commercially available facial feature tracking system for the real time analysis of facial and head movements. Mental states are inferred from the facial and head movements with a multilevel modeling approach with Dynamic Decision Networks. The individual facial and head movements are at the lowest level of the hierarchy. These movements are then used to recognize Action Units (AU) à la Ekman's Facial Action Coding System [Ekman and Friesen 1978]. For example, AU 4 (brow lowerer) and AU 7 (lid tightener) are two action units that are commonly associated with confusion [McDaniel, D'Mello, King, Chipman, Tapp and Graesser 2007]. Displays, or combinations of AUs, occupy the next level of this multilevel hierarchy. Example displays include mouth open (combination of AUs 25 and 26 or jaw drop and lips part, respectively), and head nod (combination of AUs 53 and 54 or head up and head down, respectively). The displays are then combined for mental state inference. In this fashion, the multilevel model incorporates different levels of spatial and temporal detail in a hierarchical manner that was inspired by models of how humans perceive facial activity.

Multimodal affect detection. The system uses a decision-level fusion algorithm where each channel (conversational cues, face, posture) independently provides its own diagnosis of the student's affective state. These individual diagnoses are combined with an algorithm that selects a single affective state and a confidence value of the detection. The algorithm relies on a voting rule enhanced with a few simple heuristics.

A spreading activation network is used to model decision-level fusion [Rumelhart et al. 1986]. The network consists of emotions and sensors as nodes with *projecting* and *lateral* links among them. Each sensor node is connected to each emotion node in the network by a projecting link. The degree to which a particular sensor activates a particular emotion is based on the accuracy by which the sensor has detected the emotion in past offline evaluations. Hence, if one sensor is more accurate at detecting boredom than confusion, it will excite the boredom node more than the confusion node, even if its current estimates on the probability of both emotions are approximately equivalent.

It is also possible to connect each to every other emotion with an excitatory or inhibitory *lateral* link. Related emotions excite each other while unrelated emotions inhibit each other. For example, confusion would excite frustration but boredom would inhibit engagement. It should be noted that these lateral links are sometimes not activated when the network is instantiated in some of our simulations. Indeed, the configuration of links and valences are varied in these simulations.

Each emotion node receives activation from both links and maintains an *activation value*. At any time, the emotion node with the highest activation value is considered to be the emotion that the student is currently experiencing. The decision-level fusion algorithm operates in four phases.

1. *Detection by Sensors*. Each sensor provides an independent estimate of the likelihood that the student is experiencing an emotion. The likelihood can be represented as a probability value for each emotion (e.g., the posture sensor expresses a .53 probability that the current emotion is boredom).
2. *Activation from Sensors*. Sensors spread activation and emotion nodes aggregate this activation.
3. *Activation from Emotion Nodes*. Each emotion spreads the activation received from the sensors to the other emotions, so that some emotions are excited while others are inhibited.
4. *Decision*. The emotion with the highest activation is selected to be the emotion that the student is currently experiencing.

5.3 Regulating Negative Affective States

Despite the complexity associated with real-time affect detection, detection is only one piece of the puzzle. The next challenge is to help students regulate their affective states so that positive states such as flow/engagement and curiosity persevere, while negative states such as frustration and boredom are prevented or regulated when they arise. As an initial step, we focused on interventions that help students regulate the negative affective states of boredom, frustration, and confusion.

Foundations of affect-sensitivity. An examination of the education and tutoring literature did not provide any clear guidance on how to best respond to students' affective states during tutoring. Some theories did address the presence of certain negative affective states, so insights gleaned from these theories were applied to respond to boredom, frustration, and confusion. The major theories that were considered included attribution theory [Batson et al. 1995; Heider 1958; Weiner 1986], empathy [Dweck 2002; Lepper and Chabay 1988], cognitive disequilibrium during learning [Festinger 1957; Graesser and Olde 2003; Piaget 1952], and politeness [Brown and Levinson 1987; Wang et al. 2008]. A discussion of how these theories informed affect-sensitive responses appears in [D'Mello, Craig, Fike and Graesser 2009].

In addition to the theoretical considerations, the assistance of experts in tutoring was enlisted to help create the set of affect-sensitive tutor responses. When there was no guidance from theory or expertise, the research group added affect-sensitive rules that were intuitively plausible. So in a nutshell, the rules were determined by theory, experts, and intuition. As such, the specific rules that were developed should be considered to be an initial attempt towards achieving affect-sensitivity in a tutoring context rather than a rigid set of specifications. The architecture of the Affective AutoTutor (discussed in Section 5.4) was designed to afford easy implementation and testing of a variety of responses, so the system itself can be considered to be a flexible framework to experiment with different affect-sensitive tutorial strategies.

The affect-sensitive production rules that we developed were designed to map dynamic assessments of the students' cognitive and affective states with appropriate tutor actions. In particular, at any given turn Affective AutoTutor keeps track of five major informational parameters that provide the foundations for affect sensitivity (three affective parameters and two cognitive parameters). The three affective parameters include the current affective state detected, the confidence level of that affect classification, and the previous affective state detected. The cognitive parameters include a global measure of student ability (dynamically updated throughout the session) and the conceptual quality of the student's immediate response (see Section 3.3).

Taking these five parameters as input, Affective AutoTutor is equipped with a set of production rules to map the input parameters with appropriate tutor actions. Affective AutoTutor responds with (a) feedback for the current answer with an affective facial expression, (b) an affective statement accompanied by a matching emotional facial and vocal expression by the tutor, and (c) the next dialogue move to advance the conversation.

Feedback with affective facial expression. AutoTutor provides short feedback to each student response. There are five levels of feedback: positive, neutral-positive, neutral, neutral-negative, and negative. Each feedback category has a set of predefined phrases that the tutor randomly selects from. "Good job" and "Well done" are examples of positive feedback, while "That is not right" and "You are on the wrong track" are examples of negative feedback.

In addition to articulating the textual content of the feedback, Affective AutoTutor also modulates its facial expressions and speech prosody. Positive feedback is delivered with an *approval* expression (big smile and big nod). Neutral positive feedback receives a *mild approval* expression (small smile and slight nod). Negative feedback is delivered with a *disapproval* expression (slight frown and head shake), while the tutor makes a *skeptical* face when delivering neutral-negative feedback (see Figure 4). Neutral feedback is not accompanied by a facial expression.

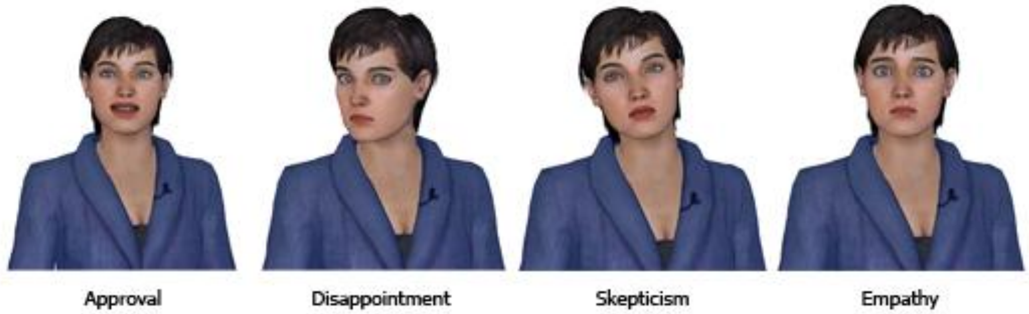


Figure 4. Synthesized facial expressions by the AutoTutor agent

Affective response. After delivering the feedback, Affective AutoTutor delivers an emotional statement if it senses that the student is bored, confused, or frustrated. A non-emotional discourse marker (e.g. "Moving on", "Try this one") is selected if the student is neutral. We have currently implemented two pedagogically distinct variants of Affective AutoTutor. These include a Supportive and a Shakeup AutoTutor.

Supportive AutoTutor. The Supportive AutoTutor responds to the students' affective states via empathetic and motivational responses. These responses always attribute the source of the students' emotion to the material instead of the students themselves. So the Supportive AutoTutor might respond to mild boredom with, "This stuff can be kind of dull sometimes, so I'm gonna try and help you get through it. Let's go". A more encouraging response is required for severe boredom (e.g., "Let's keep going, so we can move on to something more exciting"). An important point to note is that the Supportive AutoTutor never attributes the boredom to the student. Instead, it always blames itself or the material.

A response to confusion would include attributing the source of confusion to the material (e.g., "Some of this *material* can be confusing. Just keep going and I am sure you will get it") or the tutor itself (e.g., "I know I do not always convey things clearly. I am always happy to repeat myself if you need it. Try this one"). If the level of confusion is low or mild, then the pattern of responses entails: (a) acknowledging the confusion, (b) attributing it to the material or tutor, and (c) keeping the dialogue moving forward via hints, prompts, etc. In cases of severe confusion, an encouraging statement is included as well.

Similarly, frustration receives responses that attribute the source of the frustration to the material or the tutor coupled with an empathetic or encouraging statement. Examples include: "I may not be perfect, but I'm only human, right? Anyway, let's keep going and try to finish up this problem.", and "I know this *material* can be difficult, but I think you can do it, so let's see if we can get through the rest of this problem."

As a complete example, consider a student that has been performing well overall (high global ability), but the most recent contribution was not very good. If the current emotion was classified as boredom, with a high probability, and the previous emotion was classified as frustration, then AutoTutor might say the following: "Maybe this *topic* is getting old. I'll help you finish so we can try something new". This is a randomly chosen phrase from a list that was designed to indirectly address the student's boredom and to try to shift the topic before the student becomes disengaged from the learning experience. This rule fires on several different occasions, and each time it is activated the Supportive AutoTutor will select a dialogue move from a list of associated moves. In this fashion, the rules are context sensitive and are dynamically adaptive to each individual student.

Shakeup AutoTutor. The major difference between the Shakeup AutoTutor and the Supportive AutoTutor lies in the source of emotion attribution. While the Supportive AutoTutor attributes the students' negative emotions to the material or itself, the Shakeup AutoTutor directly attributes the emotions to the students. For example, possible shakeup responses to confusion are, "This material has got *you* confused, but I think you have the right idea. Try this…" and "*You* are not as confused as you might think. I'm actually kind of impressed. Keep it up".

Another difference between the two versions lies in the conservational style. While the Supportive AutoTutor is subdued and formal, the Shakeup tutor is edgier, flaunts social norms, and is witty. For example, a supportive response to boredom would be "Hang in there a bit longer. Things are about to get interesting." The shakeup counterpart of this response is "Geez this stuff sucks. I'd be bored too, but I gotta teach what they tell me".

Affective facial expression and affectively modulated speech. The affective response is accompanied by an emotional facial expression and emotionally modulated speech. These affective expressions include empathy, mild enthusiasm, high enthusiasm, skepticism, and neutral in some cases. The facial expressions in each display were informed by Ekman's work on the facial correlates of emotion expression [Ekman and Friesen 1978].

The facial expressions of emotion displayed by Affective AutoTutor are augmented with emotionally expressive speech synthesized by the agent. The emotional expressivity is obtained by variations in pitch, speech rate, and other prosodic features. Previous research has led us to conceptualize AutoTutor's affective speech on the indices of pitch range, pitch level, and speech rate [Johnstone and Scherer 2000]. The current quality of the emotionally-modulated speech is acceptable, although there is the potential for improvement.

Next dialogue move. Finally, AutoTutor responds with a move to advance the dialogue. In the current version of Affective AutoTutor, this dialogue move is sensitive to the student's cognitive state but not to his or her affective state (see Section 3). That is, affect-sensitivity is currently limited to the tutor's short feedback and motivational responses and but not its pedagogical dialogue moves that advance the learning (i.e., pumps, hints, prompts, assertions). Future affect-sensitive interventions will focus on the tutor's pedagogical moves as well (e.g., hinting when a student is stuck and frustrated). This adaptation would increase the bandwidth of communication and allow Affective AutoTutor to respond at a more sophisticated metacognitive level.

5.4 Architecture of Affective AutoTutor

The architecture of Affective AutoTutor is presented in Figure 5. This version is somewhat different from the AutoTutor architecture that was described in Section 3.5 because it represents a newer version of the system (v3). Affect detection occurs at the client in real time and the detected state is transmitted to the AutoTutor server. The Affective Dialogue Manager integrates the diagnosis of the student's cognitive (via the *Language Analyzer* and *Assessments* modules) and affective states in order to select an action that is sensitive to their emotions and cognitions.

Information from the physical sensors including a camera and posture sensor is collected (link 1a) and transmitted to the *Affect Detector* (2a and 2b). At the same time, contextual information is collected from the AutoTutor server (1b) and transmitted to the *Affect Detector* (2c). At the end of a conversational turn, the *Affect Detector* transmits its diagnosis of the student's affective state along with the text of the student's response to the *Hub* (4). The *Language Analyzer* then analyses the text of the response (5a) by parsing the text and detecting questions and other speech acts. The *Assessments* module compares the conceptual quality of the text based on expectations and misconceptions stored in the *Curriculum Script* and with *LSA* (5b). The *Dialogue Manager,* the *Affective Dialogue Manager*, and the *Question Asking* modules plan and generate the tutor's response to the student's cognitive and affective states (5c). The response is transmitted to the client (6), where it is rendered by the animated pedagogical agent, thereby completing one affect-responsive student-tutor turn.

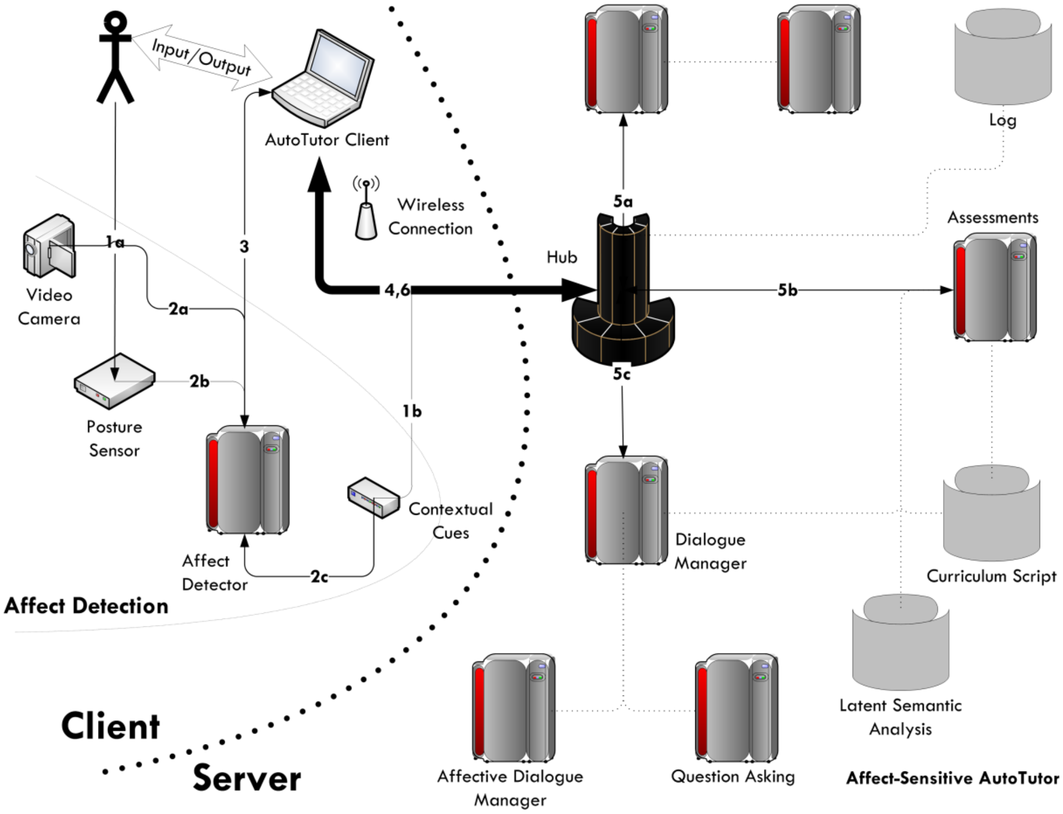


Figure 5. Architecture of Affective AutoTutor

6. EVALUATION OF AFFECTIVE AUTOTUTOR

Affective AutoTutor has been evaluated in terms of its ability to detect students' affective states and to improve learning gains when compared to the non-affective AutoTutor.

6.1 Affect Detection Accuracy

The multimodal classifier that we have developed integrates classification decisions from three unimodal classifiers: namely a dialogue-based classifier, a posture-based classifier, and a classifier that relies on facial feature tracking (See Section 5.2). We are currently in the process of systematically evaluating this classifier, so classification accuracy scores not yet available. However, we have performed extensive evaluations of earlier versions of the system [D'Mello and Graesser 2010], so the results of these evaluations are summarized here.

As a precursor to the fully automated decision-level multimodal classifier that was integrated into Affective AutoTutor, we have developed and extensively validated unimodal classifiers that focus on the dialogue, posture, and facial features and a semi-automated multimodal classifier. These classifiers were all trained and validated on data from the Multiple-Judge study described in Section 5.1. We begin with a description of the classification accuracy associated with individual channels followed by an evaluation of the semi-automated multimodal classifier.

Conversational cues (dialogue features). We compared the accuracy by which 17 standard classifiers (e.g., Naive Bayes logistic regression, support vector machines) could detect the affective states from the dialogue features [D'Mello, Craig, Witherspoon, McDaniel and Graesser 2008a]. Machine learning experiments with 10-fold cross validation indicated that standard classifiers were moderately successful in discriminating the affective states of boredom, confusion, flow/engagement frustration, and neutral, yielding a peak accuracy of 42% with neutral (*chance* = 20%) and 54% without neutral (*chance* = 25%). Individual detections of boredom, confusion, flow, and frustration, when contrasted with neutral, had accuracies of 69%, 68%, 71%, and 78%, respectively (chance = 50%). Follow-up classification analyses that assessed the degree to which machine-generated affect labels correlated with affect judgments provided by humans revealed that human-machine agreement was on par with novice judges (self and peer) but quantitatively lower than trained judges.

Gross body language (posture). As described in Section 5.2, an automated body pressure measurement system was used to capture the pressure exerted by the student on the seat and back of a chair during the tutoring session. Two algorithms to detect affect from the pressure maps were developed. The first algorithm focused on the average pressure exerted, along with the magnitude and direction of changes in the pressure during emotional experiences (see Section 5.2). The second algorithm monitored the spatial and temporal properties of naturally occurring pockets of pressure. This second algorithm was not described in this paper because it is quite complex and its performance was equivalent to the simpler algorithm.

Machine learning experiments with 10-fold cross validation yielded affect detection accuracies of 73%, 72%, 70%, 83%, and 74%, respectively (chance = 50%), in detecting boredom, confusion, delight, flow, and frustration, from neutral [D'Mello and Graesser 2009]. Accuracies involving discriminations between combinations of two, three, four, and five affective states (excluding neutral) were 71%, 55%, 46%, and 40% with chance rates being 50%, 33%, 25%, and 20%, respectively.

Facial features. Although there is considerable research on the use of facial feature tracking to detect human emotions, a vast majority of this research has focused on the basic emotions instead of the learning-centered states that are of relevance to AutoTutor [Calvo and D'Mello 2010; Zeng et al. 2009]. Therefore, we conduced some preliminary analyses to assess the possibility of detecting the learning-centered states from facial features [Craig et al. 2008; McDaniel, D'Mello, King, Chipman, Tapp and Graesser 2007]. The Mindreader system (see Section 5.2) was in development at this time because we were exploring some of the technical challenges associated with the accurate automated detection of facial expressions, which is still an open problem. Therefore, as an initial step, we had two trained judges code a sample of the observations of emotions on the action units and assessed the performance of a linear discriminant function classifier on this sample [McDaniel, D'Mello, King, Chipman, Tapp and Graesser 2007].

The classifier was also able to successfully discriminate between boredom, confusion, delight, frustration, and neutral with an accuracy of 49%, a significant improvement over the base rate (chance) of 22%. We also computed accuracy scores for individually detecting each of the affective states. The results indicated that the discriminant function was most successful in detecting delight (72%, chance = 20%) followed by confusion (58%, chance = 28%). This was expected since both these states are typically accompanied by animated facial expressions. The reliability of detecting the more subtle affective state of boredom was lower than delight and confusion (23%, chance = 12%). The results also indicated that the discriminant analysis was unable to distinguish frustration from the other emotions. In fact, the accuracy score for this emotion reflected an accuracy equal to random guessing (17%, chance = 18%).

Semi-automated multimodal affect classifier. We developed a semi-automated feature level multimodal classifier [D'Mello and Graesser 2010]. The classifier is considered to be semi-automated because while dialogue and posture features were automatically computed, facial features consisting of action units [Ekman and Friesen 1978] were manually annotated by trained humans. It is a feature-level classifier because it combines features from individual modalities and uses this combined feature vector for classification. This is different from decision-level fusion where individual classifications from the three sensors are combined via some voting rule to yield a final classification. Comparisons of feature-level and decision-level fusion on the present data set yielded equivalent performance [D'Mello 2009], so we focus on results from feature-level fusion here.

The evaluations on the unimodal classifiers discussed above used within-students cross validation methods. Therefore, different instances from the same student could be in both the training and the test sets. Individual differences play an important role in affect expression, so it is prudent to consider an evaluation method that transcends individual differences. We addressed this concern by assessing the classification accuracy of the multimodal classifier with a split-half evaluation method. Fourteen of the 28 students from the Multiple-Judge study were randomly selected and their instances were assigned to the training set. Instances from the remaining 14 students were assigned to the test set. Discriminant models were constructed from the training instances and evaluated on the testing instances.

The discriminant models yielded a 48.8% accuracy on the unseen test set for discriminating between boredom, flow/engagement, confusion, frustration, and neutral (chance = 20%). Precision (p) and recall (r) scores were: boredom (p = .39, r = .65), flow/engagement (p = .59, r = .53), confusion (p = .52, r = .49), frustration (p = .43, r = .44), and neutral (p = .38, r = .33). These results are positive because they imply that these moderate accuracy scores can be expected in real-world situations where the affect detector has to classify the emotions of unknown students.

6.2 Efficacy of Affective AutoTutor in Increasing Learning

We conducted two experiments to evaluate the pedagogical effectiveness of the two versions of Affective AutoTutor (Supportive and Shakeup as described in Section 5.3) when compared to the original tutor (Regular AutoTutor). This original AutoTutor has a conventional set of fuzzy production rules that are sensitive to cognitive states of the student, but not to the student's emotional states. Both versions of Affective AutoTutor are sensitive to the student's affective states in distinct ways. The obvious prediction is that learning gains should be superior for Affective AutoTutor when compared to the Regular AutoTutor.

Within-subjects experiment comparing Supportive, Shakeup, and Regular tutors[4]. *Participants and Design.* The experiment had a repeated-measures design where 36 undergraduate students from a university in the U.S participated in three

---

[4] This experiment is described in more detail since it has not been previously published.

computer literacy tutorial sessions on different topics, one with the Regular AutoTutor (no affect sensitivity), one with the Supportive AutoTutor, and one with Shakeup AutoTutor. For example, a student could have been assigned the hardware topic with the Supportive tutor, operating systems with the Shakeup tutor, and the Internet with the Regular tutor. The order in which students used these versions of AutoTutor and the computer literacy topics assigned to each version (i.e., hardware, operating systems, the Internet) was counterbalanced across students with a Latin Square.

Knowledge Tests. Students were tested on their knowledge of computer literacy topics both before and after the tutorial session (pretest and posttest, respectively). The testing materials were adapted from computer literacy tests used in previous experiments involving AutoTutor [Graesser, Lu, Jackson, Mitchell, Ventura, Olney and Louwerse 2004]. They were comprised of questions that assessed students' knowledge of all three computer literacy topics at deeper levels of comprehension. Each test contained 24 multiple-choice questions: 8 questions on hardware, 8 questions on operating systems, and 8 questions on Internet. Students completed alternate test versions for pretest and posttest. The two test versions, composed of different questions, tested learners on the same subject matter and content. The assignment of test versions to pretest versus posttest was counterbalanced across students.

*Procedure***.** Students were tested individually during a two and a half hour session. First, students completed an informed consent and then the pretest. Next, the general features of AutoTutor's dialogue and pedagogical strategies were described to the students. Students interacted with one of the version of AutoTutor until three main questions were successfully answered or the 30 minute training period had elapsed. They then interacted with another version of AutoTutor followed by the third version of AutoTutor. Finally, students completed the posttest and were debriefed.

Results. The dependent measures was proportional learning gains associated with each version of AutoTutor. Proportional learning gains were computed as: (posttest – pretest) / (1-pretest). We conducted a 3 × 2 ANOVA on proportional learning gains with tutor type (Regular, Supportive, Shakeup) as a within subjects factor and prior knowledge (low and high median split on pretest scores) as a between subjects factor. The analysis did not yield a significant main effect for tutor type nor a significant tutor type × prior knowledge interaction. However, there was a 0.18 sigma trend in favor of the Supportive tutor compared to the Regular tutor and a 0.28 sigma trend for the Supportive tutor over the Shakeup tutor. Learning gains for the Shakeup and Regular tutors were on par.

There is a concern pertaining to our use of a within-subjects experimental design in lieu of a between-subjects design. The obvious concern with a within-subjects design is that participating in the second session on a related subject matter might cause interference with acquired knowledge in the first session. We attempted to address this concern by only considering each student's first AutoTutor session, which was either Regular, Supportive, or Shakeup. This resulted in only 12 students in each group; a sample size too small for meaningful tests of significance but useful to explore interesting trends. The trend indicated that there was a medium effect of 0.51 sigma in favor of the Supportive Tutor when compared to the Regular tutor and the Shakeup tutor, which were similar to each other.

The within-subjects and the pseudo between-subjects analyses suggest that it is the Supportive tutor that yields the highest learning gains. Hence, we attempted to replicate this finding in a between subjects design with a larger sample of students.

Between-subjects experiment comparing Supportive and Regular tutors. We conducted an experiment to evaluate the effectiveness of the Supportive AutoTutor when compared to the Regular AutoTutor [D'Mello, Lehman, Sullins, Daigle, Combs, Vogt, Perkins and Graesser 2010d]. This experiment did not include the Shakeup AutoTutor, because the previous within-subjects experiment indicated that the Supportive Tutor consistently outperformed the Shakeup Tutor. Our prediction was that learning gains should be superior for the Supportive AutoTutor over the Regular AutoTutor.

The experiment had a between-subjects design in which 84 undergraduate students from a university in the U.S. were randomly assigned to either the Regular or the Supportive AutoTutor. Students in each condition completed a pretest on three computer literacy topics (hardware, operating systems, versus the Internet). They then completed two training sessions with the same version of AutoTutor, but on two *different* computer literacy topics. Students did not receive tutoring for the third computer literacy topic. They then completed a different posttest on all three topics. Proportional learning gains for the two topics they were tutored on served as a measure of learning, whereas gains for the topic that they did not receive tutoring on were used to assess knowledge transfer.

Students were assigned to either a low or a high prior-knowledge group based on a median split on their pretest scores. The analyses consisted of three 2 (tutor: regular vs. supportive) × 2 (prior knowledge: low vs. high) between-subjects ANOVAs for learning gains in each session and for knowledge transfer.

The tutor × prior-knowledge interaction was significant for Session 1 (see Figure 6). There was no difference in learning gains across tutors for the low prior-knowledge students (*d* = .017). However, there was a significant and substantial effect (*d* = .824) in favor of the Regular tutor for students with high prior-knowledge.

There was a different pattern in learning gains for students' second AutoTutor session. There was a significant tutor × prior-knowledge interaction. Low prior-knowledge students learned significantly more from the Supportive AutoTutor than the Regular tutor (*d* = .713). Although there was no significant difference in learning gains across tutors for the high prior knowledge students, there was a small to medium sized effect in favor of the Regular tutor (*d* = .384) (see Figure 6).

There was a nonsignificant tutor × prior-knowledge interaction for transfer scores. However, there was a medium sized effect (*d* = .583) in favor of the Supportive tutor for the low prior-knowledge students (see Figure 6). It should also be noted that the high prior-knowledge students achieved lower transfer scores than the low prior-knowledge students irrespective of whether they interacted with either the Supportive or the Regular AutoTutor. This can be attributed to the fact that AutoTutor is less effective in promoting learning for students with high domain knowledge (as mentioned in Section 4.4).

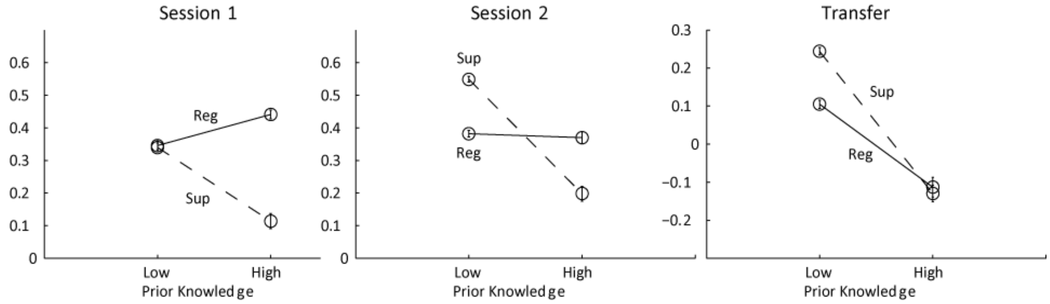


Figure 6. Interactions between prior knowledge (low vs. high) and version of AutoTutor (Regular vs. Supportive). Y-axis is proportional learning gains.

6.3 Summary of Findings, Limitations, and Future Directions

The results of this experiment support a number of conclusions regarding the effectiveness of affect-sensitivity in promoting deep learning gains. First, the Supportive AutoTutor was more effective than the Regular tutor for low-domain knowledge students in the second session, but not the first session. Therefore, it is inappropriate for the tutor to be supportive to these students before there has been enough context to show there are problems. Simply put, don't be supportive until the students need support! Second, the students with more knowledge never benefited from the Supportive AutoTutor. These students don't need that emotional support, but rather they need to go directly to the content. Hence, there appears to be a liability to quick support and empathy compared to no emotional sensitivity.

The central message is that there is an appropriate time for a particular emotion exhibited by AutoTutor. Just as there is a "time for telling" [Schwartz and Bransford 1998], there is a "time for emoting." We could imagine a trajectory where low-knowledge students start out with a non-emotional Regular tutor until they see there are problems. Then after that they need support, as manifested in Session 2 of our study. Regarding high-knowledge students, they are perfectly fine working on content for an hour or more and may get irritated with an AutoTutor showing care. But later on there may be a time when they want a Shakeup AutoTutor for stimulation, challenge, and a playful exchange. Or maybe even a Supportive AutoTutor. These are all plausible hypotheses to explore for future research.

It is also important to consider the feasibility of deploying affect detectors that require customized and often expensive hardware in classroom settings. The Affective AutoTutor used a combination of conversational cues, facial-feature tracking, and posture to detect student affect. Scalability is not a concern for the context-based affect detector because the use of conversational cues and other discourse constituents requires no additional sensors. Face-based detection is also feasible in the classroom because the only hardware requirement is a web cam However, the posture tracker (BPMS) is expensive and requires specialized hardware and software, so it is not ideal for classroom deployment. Fortunately, a number of cost-effect alternatives to the BPMS exit, thereby making real-world posture tracking viable. Some of these include pressure mats that are constructed in-house [Arroyo et al. 2009], Wii Fit Balance Boards™ [Olney and D'Mello 2010], and motion filtering techniques to assess body movements from live video [D'Mello et al. 2012]. Future research will be needed to pursue these and other alternatives, although it appears that the most scalable option would probably involve sensor-free affect detection, where predictive assessments of student affect using only context.

7. OTHER VARIANTS AND EVOLVING VERSIONS OF AUTOTUTOR

Our discussion so far has focused on the basic AutoTutor system as well as recent advancements to endow AutoTutor with affect sensing capabilities. We have also developed several other versions of AutoTutor in addition to these major research thrusts. This section briefly discusses some of these versions by distinguishing systems that have already been implemented and tested (i.e., implemented systems) from new systems that are currently in various phases of implementation and testing (evolving systems).

7.1 Variants of AutoTutor that have been Implemented and Tested

Versions of AutoTutor have been designed to incorporate particular pedagogical goals and cover different subject matters. Some of these are briefly discussed below.

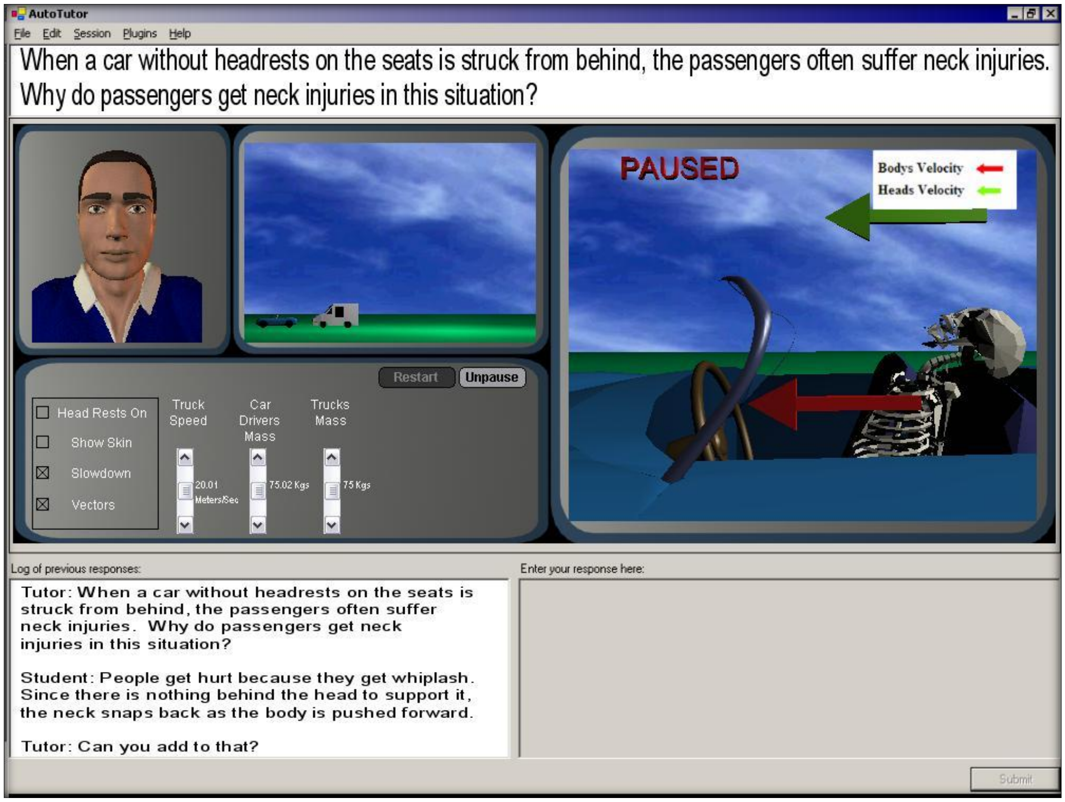
Versions that vary features of the AutoTutor agent. Most versions of AutoTutor have an animated conversational agent with synthesized speech, a small number of facial expressions, and some rudimentary hand and head gestures. These full versions have been compared to versions with voice only, text only, and various combinations of modalities in presenting AutoTutor's dialogue messages [Graesser et al. 2003]. The full animated conversational agent has shown advantages in promoting learning over alternative modalities under some conditions, particularly for deeper levels of learning [Atkinson 2002; Moreno et al. 2001]. However, available research on AutoTutor suggests that it is the verbal content of the tutor's messages that has the biggest impact on learning gains [Graesser, Moreno, Marineau, Adcock, Olney and Person 2003].

Speech-enabled AutoTutor. The traditional AutoTutor system uses synthesized speech to speak the content of its dialogue moves. However, students type their responses to the tutor's questions, thereby causing a mismatch between tutor and student input modalities. We developed a new version of AutoTutor that supported spoken input. Students spoke their responses in the new speech-enabled AutoTutor, and the commercially available Dragon Naturally Speaking ™ program was used for automatic speech recognition [D'Mello et al. 2011; D'Mello, King and Graesser 2010b].

Comparisons of the speech-enabled version of AutoTutor to the traditional typed-input version yielded significant learning independent of input modality (i.e., spoken vs. typed). There was no significant difference across modalities despite considerable automatic speech recognition errors (*word error rate* = .46) and these errors were not correlated with learning gains. The fact that performance did not degrade in light of speech recognition errors is testament of the robustness of AutoTutor's natural language processing capabilities.

AutoTutor-3D. We developed a version of AutoTutor, called AutoTutor-3D, that guides students on using interactive simulations of physics microworlds [Graesser, Chipman, Haynes and Olney 2005; Jackson and Graesser 2006]. For each of the physics problems, we developed an interactive simulation world with people, vehicles, objects, and the spatial setting associated with the problem. Figure 7 shows an example of one of these physics microworlds on a problem that involves a rear-end collision of a truck with a car. The student modifies parameters of the situation (e.g., mass of vehicles, speed of vehicles, distance between vehicles) and then asks the system to simulate what will happen. Students are also prompted to describe what they see. Their actions and descriptions are evaluated with respect to covering the expectations or matching misconceptions. AutoTutor manages the dialogue with hints and suggestions that scaffold the learning process with dialogue. The simulations were effective in increasing learning but only for those students who invested sufficient effort to run the simulations.

AutoTutor with enhanced feedback. We created different versions of AutoTutor in order to manipulate the feedback that the college students received during their interactions with the tutor [Jackson and Graesser 2007]. Students received either *content feedback* (e.g., highlighting important words after students type them in), *progress feedback* (e.g., displaying points on their performance) feedback, both, or neither. An experiment comparing the effects of these versions was quite illuminating in a number of respects. Although students learned in all feedback conditions, it was the content feedback that had a greater impact on learning than the progress feedback. We were surprised to learn that the students' perceptions of these systems were inversely related to the amount they learned. Indeed, the mean ratings of nearly all of the scales of student perceptions were more positive in those conditions that yielded the least amount of learning. Simply put, deep learning is not fun.



When a car without headrests on the seats is struck from behind, the passengers often suffer neck injuries. Why do passengers get neck injuries in this situation?

☐ Head Rests On
☐ Show Skin
☒ Slowdown
☒ Vectors

Truck Speed: 20.01 Meters/Sec
Car Drivers Mass: 75.02 Kgs
Trucks Mass: 75 Kgs

Log of previous responses:

Tutor: When a car without headrests on the seats is struck from behind, the passengers often suffer neck injuries. Why do passengers get neck injuries in this situation?

Student: People get hurt because they get whiplash. Since there is nothing behind the head to support it, the neck snaps back as the body is pushed forward.

Tutor: Can you add to that?

Enter your response here:

Figure 7. AutoTutor-3D with simulations

7.2 New and Evolving Systems based on the AutoTutor Technology

Aspects of the AutoTutor system are finding use in a number of new technologies that we and our collaborators are developing. Some of these systems are not AutoTutor per se, but are inspired by AutoTutor or incorporate some aspect of AutoTutor. These are briefly described below along with the name of the chief architect of each system who we are collaborating with.

AutoTutor-Lite (*Xiangen Hu, University of Memphis*)**.** In stark contrast to the computationally intensive AutoTutor-3D system, there is a minimalistic version of AutoTutor called AutoTutor-Lite [Hu et al. 2009]. AutoTutor-Lite includes the AutoTutor-style interface and interaction (animated agent and natural language conversation), but with a lightweight language analyzer and dialogue manager. AutoTutor-Lite has excellent authoring tools that lesson planners and instructors can use, even when they have minimal computer skills. Moreover, AutoTutor-Lite can be applied to PowerPoint content on any verbal subject matter, is easily customizable, and can be integrated into e-learning environments on the web as well as the desktop. One can imagine an industry that "autotutorizes" the conventional eLearning content that is widely available.

Operation ARIES (*Keith Millis, Northern Illinois University)* ARIES (Acquiring Research Investigative and Evaluative Skills) [Cai et al. 2009; Millis et al. 2011] teaches scientific reasoning and critical thinking skills with two animated pedagogical agents. One agent in ARIES, called the guide-agent, is an expert on scientific inquiry and serves as a knowledgeable tutor. The other agent is a fellow student that exhibits low knowledge or flawed knowledge that the guide agent and human student will hopefully correct. An imperfect agent may indeed help learning. This notion has been pursued in the Teachable Agent research of Biswas and colleagues [Biswas et al. 2005]. Human students attempt to help a fellow student agent who has misconceptions and incomplete knowledge. The process of the human student trying to help the cyber student actually ends up facilitating learning in the human. In ARIES, a case study of an experiment is presented which may or may not have a number of flaws with respect to scientific methodology. A 3-way conversation transpires (called a *trialog*ue) among the human student, the expert agent, and the student agent. ARIES facilitates learning compared to normal training methods for scientific critical thinking [Millis, Forsyth, Butler, Wallace, Graesser and Halpern 2011].

Deep Tutor (*Vasile Rus, University of Memphis*)**.** We are working on systems that analyze language and discourse at deeper levels. Researchers can move beyond LSA and regular expressions and into more structure-sensitive processing and semantic decomposition [Rus and Graesser 2007; Rus et al. 2008]. Some versions of the dialogue manager module have lists of production rules and finite-state grammars, but also move to new levels into the realm of recursive, complex planning and multiple-goal agendas. This approach of deeper natural language processing and discourse management is currently being developed and tested in the area of physics via DeepTutor

GuruTutor (*Andrew Olney, University of Memphis*)**.** AutoTutor was modeled after novice human tutors and extensions that had more ideal tutoring mechanisms. The human tutors that served as guides were untrained in tutoring skills and had moderate domain knowledge; they were peer tutors, cross-age tutors, or paraprofessionals, but rarely were accomplished professionals or expert tutors. Since there is some evidence that expert tutors are more effective in promoting learning gains than their unaccomplished counterparts [Bloom 1984; Cohen, Kulik and Kulik 1982], we are currently in the process of building a tutoring system (Guru) for high school biology based on the tactics, actions, and dialogue of expert human tutors. The pedagogical and motivational strategies of Guru are informed by a detailed computational model of expert human tutoring [D'Mello, Hays, Williams, Cade, Brown and Olney 2010a; D'Mello, Lehman and Person 2010c; D'Mello, Olney and Person 2010e]. In addition to encompassing some of the ideal tutoring strategies of the expert tutors, Guru also includes deeper natural language understanding capabilities, dynamic conversational management, and an engaging agent embedded in a 3D gaming world.

MetaTutor (*Roger Azevedo, McGill University*)**.** While AutoTutor and Affective AutoTutor focus on students' cognitive and affective processes, MetaTutor is an ITS that tracks, models, and regulates students' metacognitive processes during the learning of complex scientific content (e.g., functioning of the circulatory system) within hypermedia environments [Azevedo et al. 2008]. MetaTutor was based on extensive research documenting the importance of metacognitive processes such as monitoring, goal setting, and planning during learning [Azevedo 2009]. Unfortunately, students rarely deploy these metacognitive processes productively during unscaffolded learning with hypermedia [Azevedo et al. 2005]. Humans who are trained as external regulation agents can be quite effective in helping students deploy key metacognitive and self regulated learning processes in a manner that is contextually coupled with the learning material and environment. Therefore, MetaTutor was designed to automate the role of an externally regulating agent while students learn with hypermedia. Over time, MetaTutor is designed to fade into the background in an attempt to get students to eventually regulate their own learning.

AutoMentor (*David Shaffer, University of Wisconsin-Madison*). AutoMentor will be an automated virtual mentor that provides guidance to students as they interact in groups in serious games. AutoMentor to some extent builds on AutoTutor, however, while AutoTutor manages one-on-one tutorial dialogue, AutoMentor will interact with groups of players in the serious game. AutoMentor will be integrated in the serious game on urban planning called Land Science. The game helps students understand the kinds of problems and problem solving that socially valued professions routinely engage in. For example, how is the development of cities and suburbs influenced by zoning, roads, parks, housing, and economic investment? What developments in science and high-quality information sources need to be communicated in justifications of decisions? Currently, a human mentor can broadcast suggestions to the student players simultaneously whenever the mentor deems it appropriate to send a message. The idea is to replace the suggestions generated by the human with AutoMentor's suggestions.

8. CONCLUDING REMARKS

We presented AutoTutor, Affective AutoTutor, and existing and evolving versions of these systems as educational technologies that exemplify some of the exciting technologies emerging from the Intelligent Interactive Systems community. These 21st century systems offer considerable advantages over the 20th century computer-based training systems. They are also improvements over many of the 1990s ITSs that impressively modeled students' cognitive states, while largely ignoring motivation and emotion. The ability to yield impressive learning gains that match human tutors by modeling and responding to students' cognitive, emotional, and motivational processes is a major contribution of these systems. Additional refinements of these systems, as well as some of the newer technologies, such as AutoTutor-Lite, GuruTutor, DeepTutor, ARIES, MetaTutor, and AutoMentor, are expected to yield even greater improvements in engagement, self-efficacy, task persistence, and motivation.

Another impressive achievement of these systems, particularly the dialogue-based tutors, is that they are redefining the human-computer interaction paradigm. Most (but not all) of the 20th century systems required humans to communicate with computers through windows, icons, menus, and pointing devices (i.e., the WIMP paradigm). But humans have always communicated with each other through speech and a host of non-verbal cues such as facial expressions, paralinguistic features of speech, oculesics (eye contact), posture, and gesture. In addition to enhancing the content of the message, these communicative channels provide information regarding the cognitive states, affective states, motivation levels, and social dynamics of the communicators. In their endeavor to support speech input, natural language dialogues, affect sensing, gaze tracking, and embodied conversational agents, the systems developed by us and our colleagues are pushing the envelope by blurring the boundary between human-human and human-computer interactions.

Although the enhanced interactivity of the 21st century systems is impressive, it is always important to realize that content reigns supreme for learning. When it comes to deep learning, the medium is not the message, but the message is the message. Therefore, advances in intelligence and interactivity of educational technologies should always be accompanied by advances in the content-based learning sciences.

Just as it takes a village to raise a child, it takes an interdisciplinary milieu of researchers encompassing computer science, engineering, psychology, linguistics, and education to develop many of the advanced learning technologies including but not limited to AutoTutor and Affective AutoTutor. And it will take even broader teams of interdisciplinary minds to position us towards the next set of discoveries and innovations that will undoubtedly emerge from the field of interactive intelligent systems over the next decade.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

The authors would like to thank members of the Tutoring Research Group (http://www.autotutor.org/) and the Memphis Affective Computing Group (http://emotion.autotutor.org/) at the University of Memphis. We gratefully acknowledge our collaborators including Keith Millis (Northern Illinois University), Andrew Olney (University of Memphis), Rosalind Picard (MIT), Natalie Person (Rhodes College), Vasile Rus (University of Memphis), David Shaffer (University of Wisconsin), and Xiangen Hu (University of Memphis).

REFERENCES

AFZAL, S. AND ROBINSON, P. 2011. Natural affect data: Collection and annotation. In *New Perspectives on Affect and Learning Technologies*, R. CALVO AND S. D'MELLO, Eds. Springer, New York, NY, 44-70.

ALEVEN, V. AND KOEDINGER, K. 2002. An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26, 2, 147-179.

ALEVEN, V., MCLAREN, B., ROLL, I. AND KOEDINGER, K. 2006. Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*,16, 2, 101-128.

ANDERSON, J., CORBETT, A., KOEDINGER, K. AND PELLETIER, R. 1995. Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*,4, 2, 167-207.

ANDERSON, J., DOUGLASS, S. AND QIN, Y. 2005. How should a theory of learning and cognition inform instruction? In *Experimental Cognitive Psychology and its Applications*, A. HEALY, Ed. American Psychological Association, Washington, DC., 47-58.

ANDERSON, J.R. AND GLUCK, K. 2001. What role do cognitive architectures play in intelligent tutoring systems? In *Cognition & Instruction: Twenty-five Years of Progress*, D. KLAHR AND S.M. CARVER, Eds. Erlbaum, Hillsdale, NJ, 227-262.

ARROYO, I., WOOLF, B., COOPER, D., BURLESON, W., MULDNER, K. AND CHRISTOPHERSON, R. 2009. Emotion sensors go to school. In *Proceedings of the 14th International Conference on Artificial Intelligence In Education*, V. DIMITROVA, R. MIZOGUCHI, B. DU BOULAY AND A. GRAESSER, Eds. IOS Press, Amsterdam, 17-24.

ATKINSON, R. 2002. Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, 94, 2, 416-427.

AZEVEDO, R. 2009. Theoretical, methodological, and analytical challenges in the research on metacognition and self-regulation: A commentary. *Metacognition & Learning*, 4, 87-95.

AZEVEDO, R., CROMLEY, J., WINTERS, F., MOOS, D. AND GREENE, J. 2005. Adaptive human scaffolding facilitates adolescents' self-regulated learning with hypermedia. *Instructional Science*, 33, 5-6, 381-412.

AZEVEDO, R., WITHERSPOON, A., GRAESSER, A., MCNAMARA, D., RUS, V., CAI, Z., LINTEAN, M. AND SILER, E. 2008. MetaTutor: An adaptive hypermedia system for training and fostering self-regulated learning about complex science topics. In *Papers from the Association for the Advancement of Artificial Intelligence Fall Symposium on Cognitive and Metacognitive Educational Systems*, R. PIRRONE, R. AZEVEDO AND G. BISWAS, Eds. AAAI Press, Menlo Park, CA, 14-19.

BAKER, R., D'MELLO, S., RODRIGO, M. AND GRAESSER, A. 2010. Better to be frustrated than bored: The incidence and persistence of affect during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68, 4, 223-241.

BAKER, R. AND YACEF, K. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1, 1, 3-17.

BATSON, C., TURK, C., SHAW, L. AND KLEIN, T. 1995. Information function of empathic emotion - learning that we value the others welfare. *Journal of Personality and Social Psychology*,68, 2, 300-313.

BISWAS, G., LEELAWONG, K., SCHWARTZ, D. AND VYE, N. 2005. Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, 19, 3-4, 363-392.

BLOOM, B. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain.* McKay, New York.

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.

BOEKAERTS, M. 2007. Understanding students' affective processes in the classroom. In *Emotion in Education*, P. SCHUTZ AND R. PEKRUN, Eds. Academic Press, San Diego, CA, 37-56.

BRANSFORD, J., GOLDMAN, S. AND VYE, N. 1991. Making a difference in people's ability to think: Reflections on a decade of work and some hopes for the future. In *Influences on Children*, R. STERNBERG AND L. OKAGAKI, Eds. Erlbaum, Hillsdale, NJ, 147-180.

BROWN, A., ELLERY, S. AND CAMPIONE, J. 1998. Creating zones of proximal development electronically. In *Thinking Practices in Mathematics and Science Learning*, J. GREENO AND S. GOLDMAN, Eds. Lawrence Erlbaum, Mahwah, NJ, 341-367.

BROWN, P. AND LEVINSON, S. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge.

BULL, P. 1987. *Posture and Gesture*. Pergamon Press, Oxford

BURLESON, W. AND PICARD, R. 2007. Evidence for gender specific approaches to the development of emotionally intelligent learning companions. *IEEE Intelligent Systems*,22, 4, 62-69.

CAI, Z., GRAESSER, A.C., MILLIS, K.K., HALPERN, D., WALLACE, P., MOLDOVAN, C. AND FORSYTH, C. 2009. ARIES!: An intelligent tutoring system assisted by conversational agents. In *Artificial Intelligence in Education - Building Learning Systems That Care: from Knowledge Representation to Affective Modelling*, V. DIMITROVA, R. MIZOGUCHI, B. DUBOULAY AND A. GRAESSER, Eds. IOS Press, Amsterdam, 796.

CALVO, R. AND D'MELLO, S. 2011. *New Perspectives on Affect and Learning Technologies*. Springer, New York.

CALVO, R.A. AND D'MELLO, S.K. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*,1, 1, 18-37.

CHAFFAR, S., DERBALI, L. AND FRASSON, C. 2009. Inducing positive emotional state in intelligent tutoring systems. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, V. DIMITROVA, R. MIZOGUCHI, B. DU BOULAY AND A. GRAESSER, Eds. IOS Press, Amsterdam, 716-718.

CHI, M., DELEEUW, N., CHIU, M. AND LAVANCHER, C. 1994. Eliciting self-explanations improves understanding. *Cognitive Science*,18, 3, 439-477.

CHI, M., ROY, M. AND HAUSMANN, R. 2008. Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science*,32, 2, 301-341.

CHI, M., SILER, S. AND JEONG, H. 2004. Can tutors monitor students' understanding accurately? *Cognition and Instruction*,22, 3, 363-387.

CHI, M., SILER, S., JEONG, H., YAMAUCHI, T. AND HAUSMANN, R. 2001. Learning from human tutoring. *Cognitive Science*,25, 4, 471-533.

COHEN, J. 1992. A power primer. *Psychological Bulletin*,112, 1, 155-159.

COHEN, P., KULIK, J. AND KULIK, C. 1982. Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*,19, 2, 237-248.

CONATI, C. AND MACLAREN, H. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*,19, 3, 267-303.

CONATI, C., MARSELLA, S. AND PAIVA, A. 2005. Affective interactions: The computer in the affective loop. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, J. RIEDL AND A. JAMESON, Eds. ACM, New York, 7.

CORBETT, A. 2001. Cognitive computer tutors: Solving the two-sigma problem. In *Proceedings of the 8th International Conference on User Modeling*, M. BAUER, P. GMYTRASIEWICZ AND J. VASSILEVA, Eds. Springer, Berlin / Heidelberg, 137-147.

CORBETT, A. 2002. Cognitive tutor algebra I: Adaptive student modeling in widespread classroom use. In *Technology and Assessment: Thinking Ahead - Proceedings from a Workshop,* National Academy Press, Washington, D.C., 50-62.

CORBETT, A. AND ANDERSON, J. 1994. Knowledge Tracing - Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*,4, 4, 253-278.

CORBETT, A., ANDERSON, J., GRAESSER, A., KOEDINGER, K. AND VANLEHN, K. 1999. Third generation computer tutors: Learn from or ignore human tutors? In *Proceedings of Association for Computing Machinery Special Interest Group on Computer Human Interaction Conference on Human Factors in Computing Systems* ACM, Pittsburgh, PA, 85 - 86.

CRAIG, S., D'MELLO, S., WITHERSPOON, A. AND GRAESSER, A. 2008. Emote aloud during learning with AutoTutor: Applying the facial action coding system to cognitive-affective states during learning. *Cognition & Emotion*,22, 5, 777-788.

CRAIG, S., GRAESSER, A., SULLINS, J. AND GHOLSON, J. 2004. Affect and learning: An exploratory look into the role of affect in learning. *Journal of Educational Media*,29, 241-250.

D'MELLO, S. 2009. *Affect Measurement by Man and Machine* (Doctoral Dissertation), University of Memphis.

D'MELLO, S. in review. A meta-analysis on the incidence of emotions during complex learning.

D'MELLO, S., CRAIG, S., SULLINS, J. AND GRAESSER, A. 2006. Predicting affective states expressed through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education*,16, 1, 3-28.

D'MELLO, S., CRAIG, S., WITHERSPOON, A., MCDANIEL, B. AND GRAESSER, A. 2008a. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*,18, 1-2, 45-80.

D'MELLO, S., DOWELL, N. AND GRAESSER, A. 2011. Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology: Applied*,17, 1, 1-17.

D'MELLO, S. AND GRAESSER, A. 2009. Automatic detection of learners' affect from gross body language. *Applied Artificial Intelligence*,23, 2, 123 - 150.

D'MELLO, S. AND GRAESSER, A. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-adapted Interaction*,20, 2, 147-187.

D'MELLO, S., HAYS, P., WILLIAMS, C., CADE, W., BROWN, J. AND OLNEY, A. 2010a. Collaborative Lecturing by Human and Computer Tutors In *Proceedings of 10th International Conference on Intelligent Tutoring Systems*, J. KAY AND V. ALEVEN, Eds. Springer, Berlin / Heidelberg, 609-618.

D'MELLO, S., JACKSON, G., CRAIG, S., MORGAN, B., CHIPMAN, P., WHITE, H., PERSON, N., KORT, B., EL KALIOUBY, R., PICARD, R. AND GRAESSER, A. 2008b. AutoTutor detects and responds to learners affective and cognitive states. In *Proceedings of the Workshop on Emotional and Cognitive Issues in Intelligent Tutoring Systems held in conjunction with the Ninth International Conference on Intelligent Tutoring Systems*, Montreal, Canada.

D'MELLO, S., KING, B. AND GRAESSER, A. 2010b. Towards spoken human-computer tutorial dialogues. *Human-Computer Interaction*,25, 4, 289-323.

D'MELLO, S., LEHMAN, B. AND PERSON, N. 2010c. Expert tutors feedback is immediate, direct, and discriminating. In *Proceedings of the 23rd Florida Artificial Intelligence Research Society Conference*, C. MURRAY AND H. GUESGEN, Eds. AAAI Press, Menlo Park, CA, 595-560.

D'MELLO, S., LEHMAN, B., SULLINS, J., DAIGLE, R., COMBS, R., VOGT, K., PERKINS, L. AND GRAESSER, A. 2010d. 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*, J. KAY AND V. ALEVEN, Eds. Springer, Berlin / Heidelberg, 245-254.

D'MELLO, S., OLNEY, A. AND PERSON, N. 2010e. Mining collaborative patterns in tutorial dialogues. *Journal of Educational Data Mining*,2, 1, 1-37.

D'MELLO, S., PICARD, R. AND GRAESSER, A. 2007. Towards an affect-sensitive AutoTutor. *IEEE Intelligent Systems*, 22, 4, 53-61.

D'MELLO, S., CHIPMAN, P. AND GRAESSER, A. 2007. Posture as a predictor of learner's affective engagement. In *Proceedings of the 29th Annual Cognitive Science Society*, D. MCNAMARA AND G. TRAFTON, Eds. Cognitive Science Society, Austin, TX, 905-991.

D'MELLO, S., CRAIG, S., FIKE, K. AND GRAESSER, A. 2009. Responding to learners' cognitive-affective states with supportive and shakeup dialogues. In *Proceedings of the 13th International Conference on Human-Computer Interaction. Part III: Ubiquitous and Intelligent Interaction*, J. JACKO, Ed. Springer, Berlin/Heidelberg, 595-604.

D'MELLO, S., DALE, R. AND GRAESSER, A. 2012. Disequilibrium in the mind, disharmony in the body. *Cognition & Emotion*,26, 2, 362-374.

D'MELLO, S. AND GRAESSER, A. 2012. Emotions during learning with AutoTutor. In *Adaptive Technologies for Training and Education*, P. DURLACH AND A. LESGOLD, Eds. Cambridge University Press, New York, NY, 117-139.

DEL SOLDATO, T. AND DU BOULAY, B. 1995. Implementation of motivational tactics in tutoring systems. *International Journal of Artificial Intelligence in Education*,6, 337-378.

DODDS, P. AND FLETCHER, J. 2004. Opportunities for new "smart" learning environments enabled by next-generation web capabilities. *Journal of Educational Multimedia and Hypermedia*,13, 4, 391-404.

DOIGNON, J.P. AND FALMAGNE, J.C. 1998. *Knowledge spaces*. Springer-Verlag New York.

DU BOULAY, B. 2011. Towards a motivationally-intelligent pedagogy: How should an intelligent tutor respond to the unmotivated or the demotivated? In *New Perspectives on Affect and Learning Technologies*, R. CALVO AND S. D'MELLO, Eds. Springer, New York, 41-52.

DWECK, C. 2002. Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways). In *Improving Academic Achievement: Impact of Psychological Factors on Education* J. ARONSON, Ed. Academic Press, Orlando, FL, 61-87.

EKMAN, P. 1992. An argument for basic emotions. *Cognition & Emotion*,6, 3-4, 169-200.

EKMAN, P. AND FRIESEN, W. 1978. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA.

EL KALIOUBY, R. AND ROBINSON, P. 2005a. Generalization of a vision-based computational model of mind-reading. In *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction*, J. TAO AND R. PICARD, Eds. Springer-Verlag Berlin-Heidelberg, 582-589.

EL KALIOUBY, R. AND ROBINSON, P. 2005b. Real-time inference of complex mental states from facial expressions and head gestures. In *Proceedings of the 2004 Conference on Computer Vision and Pattern*

Recognition Workshop B. KISACANIN, V. PAVLOVIC AND T. HUANG, Eds. Spring-Verlag, Berlin, 181–200.

EVENS, M. AND MICHAEL, J. 2006. *One-on-one Tutoring by Humans and Machines*. Erlbaum, Mahwah, NJ.

FESTINGER, L. 1957. *A Theory of Cognitive Dissonance*. Stanford University Press, Stanford, CA.

FLETCHER, J. 2003. Evidence for learning from technology-assisted instruction. In *Technology Applications in Education: A Learning View*, J. H. F. O'NEIL AND R. PEREZ, Eds. Erlbaum, Hillsdale, NJ, 79-99.

FORBES-RILEY, K., ROTARU, M. AND LITMAN, D. 2008. The relative impact of student affect on performance models in a spoken dialogue tutoring system. *User Modeling and User-Adapted Interaction*,18, 1-2, 11-43.

FOX, B. 1993. *The Human Tutorial Dialogue Project*. Lawrence Erlbaum Associates, Hillsdale, NJ.

GAGNE, R.M. 1985. *The Conditions of Learning and Theory of Instruction*. Holt, Rinehart, & Winston, New York.

GERTNER, A. AND VANLEHN, K. 2000. Andes: A coached problem solving environment for physics. In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, G. GAUTHIER, C. FRASSON AND K. VANLEHN Eds. Springer, Berlin / Heidelberg, 133-142.

GOLEMAN, D. 1995. *Emotional intelligence*. Bantam Books, New York.

GRAESSER, A., CHIPMAN, P., HAYNES, B. AND OLNEY, A. 2005. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*,48, 4, 612-618.

GRAESSER, A., CHIPMAN, P., KING, B., MCDANIEL, B. AND D'MELLO, S. 2007a. Emotions and learning with AutoTutor. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education* R. LUCKIN, K. KOEDINGER AND J. GREER, Eds. IOS Press, Amsterdam, 569-571.

GRAESSER, A., D'MELLO, S. AND PERSON, N. 2009. Meta-knowledge in tutoring. In *Metacognition in Educational Theory and Practice* D. HACKER, J. DUNLOSKY AND A.C. GRAESSER, Eds. Erlbaum, Mahwah, NJ., 361-382.

GRAESSER, A., LU, S.L., JACKSON, G., MITCHELL, H., VENTURA, M., OLNEY, A. AND LOUWERSE, M. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*,36, 180-193.

GRAESSER, A., MCDANIEL, B., CHIPMAN, P., WITHERSPOON, A., D'MELLO, S. AND GHOLSON, B. 2006. Detection of emotions during learning with AutoTutor. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, R. SUN AND N. MIYAKE, Eds. Cognitive Science Society, Austin, TX, 285-290.

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*, U. HOPPE, F. VERDEJO AND J. KAY, Eds. IOS Press, Amsterdam, 47-54.

GRAESSER, A. AND OLDE, B. 2003. How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*,95, 3, 524-536.

GRAESSER, A., PENUMATSA, P., VENTURA, M., CAI, Z. AND HU, X. 2007b. Using LSA in AutoTutor: Learning through mixed-initiative dialogue in natural language. In *Handbook of Latent Semantic Analysis*, T. LANDAUER, D. MCNAMARA, S. DENNIS AND W. KINTSCH, Eds. Erlbaum, Mahwah, NJ, 243-262.

GRAESSER, A. AND PERSON, N. 1994. Question asking during tutoring. *American Education Research Journal*,31, 1, 104-137.

GRAESSER, A., PERSON, N. AND HARTER, D. 2001a. Teaching tactics and dialogue In AutoTutor. *International Journal of Artificial Intelligence in Education*,12, 257-279.

GRAESSER, A., PERSON, N. AND MAGLIANO, J. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*,9, 6, 495-522.

GRAESSER, A., VANLEHN, K., ROSE, C.P., JORDAN, P.W. AND HARTER, D. 2001b. Intelligent tutoring systems with conversational dialogue. *AI Magazine*,22, 4, 39-51.

GRAESSER, A., WIEMER-HASTINGS, K., WIEMER-HASTINGS, P. AND KREUZ, R. 1999. AutoTutor: A simulation of a human tutor. *Cognitive Systems Research*,1, 1, 35-51.

GRAESSER, A., WIEMER-HASTINGS, P., WIEMER-HASTINGS, K., HARTER, D. AND PERSON, N. 2000. Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*,8, 2, 129-147.

HEIDER, F. 1958. *The Psychology of Interpersonal Relations*. John Wiley & Sons, New York.

HESTENES, D., WELLS, M. AND SWACKHAMER, G. 1992. Force concept inventory. *The Physics Teacher*,30, 141-158.

HU, X., CAI, Z., HAN, L., CRAIG, S.D., WANG, T. AND GRAESSER, A.C. 2009. AutoTutor Lite. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, V. DIMITROVA, R. MIZOGUCHI, B. DU BOULAY AND A. GRAESSER, Eds. IOS Press, Amsterdam, 802-802.

JACKSON, G. AND GRAESSER, A. 2006. Applications of human tutorial dialog in AutoTutor: An intelligent tutoring system. *Revista Signos*,39, 60, 31-48.

JACKSON, G.T. AND GRAESSER, A.C. 2007. Content matters: An investigation of feedback categories within an ITS. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, R. LUCKIN, K. KOEDINGER AND J.J. GREEN, Eds. IOS Press, Amsterdam, 127–134.

JOHNSON, W. AND VALENTE, L. 2008. Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. In *Proceedings of the 20th National Artificial Intelligence Conference*, M. GOKER AND K. HAIG Eds. AAAI Press, Menlo Park, CA, 1632-1639.

JOHNSTONE, T. AND SCHERER, K. 2000. Vocal communication of emotion. In *Handbook of Emotions*, M. LEWIS AND J. HAVILAND-JONES, Eds. Guilford Press, New York, 220-235.

KOEDINGER, K., ANDERSON, J., HADLEY, W. AND MARK, M. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*,8, 30-43.

KOEDINGER, K. AND CORBETT, A. 2006. Cognitive tutors: Technology bringing learning sciences to the classroom. In *The Cambridge Handbook of the Learning Sciences*, R.K. SAWYER, Ed. Cambridge University Press, New York, NY, 61-78.

LANDAUER, T., MCNAMARA, D., DENNIS, S. AND KINTSCH, W. Eds. 2007. *Handbook of Latent Semantic Analysis*. Erlbaum, Mahwah, NJ.

LEPPER, M. AND CHABAY, R. 1988. Socializing the intelligent tutor: Bringing empathy to computer tutors. In *Learning Issues for Intelligent Tutoring Systems*, H. MANDL AND A. LESGOLD, Eds. Erlbaum, Hillsdale, NJ, 242-257.

LEPPER, M. AND WOOLVERTON, M. 2002. The wisdom of practice: Lessons learned from the study of highly effective tutors. In *Improving Academic Achievement: Impact of Psychological Factors on Education*, J. ARONSON, Ed. Academic Press, Orlando, FL, 135-158.

LESGOLD, A., LAJOIE, S., BUNZO, M. AND EGGAN, G. 1992. SHERLOCK: A coached practice environment for an electronics troubleshooting job. In *Computer-Assisted Instruction and Intelligent Tutoring Systems* J.H. LARKIN AND R.W. CHABAY, Eds. Erlbaum, Hillsdale, NJ, 201-238.

LITMAN, D., ROSE, C., FORBES-RILEY, K., VANLEHN, K., BHEMBE, D. AND SILLIMAN, S. 2006. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence In Education*,16, 2, 145-170.

LITMAN, D. AND SILLIMAN, S. 2004. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, MA.

MCDANIEL, B., D'MELLO, S., KING, B., CHIPMAN, P., TAPP, K. AND GRAESSER, A. 2007. Facial features for affective state detection in learning environments. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, D. MCNAMARA AND G. TRAFTON, Eds. Cognitive Science Society, Austin, TX, 467-472.

MEYER, D. AND TURNER, J. 2006. Re-conceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review*,18, 4, 377-390.

MILLIS, K., FORSYTH, C., BUTLER, H., WALLACE, P., GRAESSER, A. AND HALPERN, D. in press. Operation ARIES! A serious game for teaching scientific inquiry. In *Serious Games and Edutainment Applications*, M. MA, A. OIKONOMOU AND J. LAKHMI, Eds. Springer-Verlag, London, UK.

MITROVIC, A., MARTIN, B. AND SURAWEERA, P. 2007. Intelligent tutors for all: The constraint-based approach. *IEEE Intelligent Systems*,22, 4, 38-45.

MOORE, J. 1995. *Participating in Explanatory Dialogues*. MIT Press, Cambridge, MA.

MORENO, R., MAYER, R., SPIRES, H. AND LESTER, J. 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, 2, 177-213.

O'NEIL, H.F. AND PEREZ, R.S. Eds. 2006. *Web-based Learning: Theory, Research, and Practice*. Lawrence Erlbaum, Mahwah NJ.

OLNEY, A. AND D'MELLO, S. 2010. Interactive Event: A DIY Pressure Sensitive Chair for Intelligent Tutoring Systems. In *Proceedings of 10th International Conference on Intelligent Tutoring Systems* J. KAY AND V. ALEVEN, Eds. Springer, Berlin / Heidelberg, 456.

OLNEY, A., LOUWERSE, M., MATHEWS, E., MARINEAU, J., HITE-MITCHELL, H. AND GRAESSER, A. 2003. Utterance classification in AutoTutor. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Canada.

PEKRUN, R., GOETZ, T., DANIELS, L., STUPNISKY, R.H. AND PERRY, R. 2010. Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*,102, 3, 531-549.

PEKRUN, R. AND STEPHENS, E.J. 2012. Academic emotions. In *American Psychological Association Educational Psychology Handbook, Vol 2: Individual Differences and Cultural and Contextual Factors*, K. HARRIS, S. GRAHAM, T. URDAN, S. GRAHAM, J. ROYER AND M. ZEIDNER, Eds. American Psychological Association, Washington, DC, 3-31.

PERSON, N. AND GRAESSER, A. 2002. Human or computer? AutoTutor, in a Bystander Turing Test. In *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, S. CERRI, G. G. AND P. F., Eds. Springer, Berlin / Heidelberg, 821-830.

PERSON, N., GRAESSER, A., MAGLIANO, J. AND KREUZ, R. 1994. Inferring what the student knows in one-to-one tutoring: The role of student questions and answers. *Learning and Individual Differences*,6, 205-229.

PERSON, N.K., GRAESSER, A.C., KREUZ, R.J. AND POMEROY, V. 2003. Simulating human tutor dialog moves in AutoTutor. *International Journal on Artificial Intelligence in Education*,12, 23-39.

PIAGET, J. 1952. *The Origins of Intelligence*. International University Press, New York.

PICARD, R. 1997. *Affective Computing*. MIT Press, Cambridge, Mass.

PICARD, R. 2010. Affective Computing: From Laughter to IEEE. *IEEE Transactions on Affective Computing*,1, 1, 11-17.

PON-BARRY, H., CLARK, B., SCHULTZ, K., BRATT, E.O. AND PETERS, S. 2004. Advantages of spoken language interaction in dialogue-based intelligent tutoring systems. In *Proceedings of Seventh International Conference on Intelligent Tutoring Systems*, J. LESTER, R. VICARI AND F. PARAGUACU Eds. Springer, Berlin / Heidelberg, 390-400.

POUR, P.A., HUSSEIN, S., ALZOUBI, O., D'MELLO, S. AND CALVO, R. 2010. The impact of system feedback on learners' affective and physiological states. In *Proceedings of 10th International Conference on Intelligent Tutoring Systems*, J. KAY AND V. ALEVEN, Eds. Springer-Verlag, Berlin / Heidelberg, 264-273.

PSOTKA, J., MASSEY, D. AND MUTTER, S. 1988. *Intelligent Tutoring Systems: Lessons Learned*. Lawrence Erlbaum Associates, Hillsdale, NJ.

ROBISON, J., MCQUIGGAN, S. AND LESTER, J. 2009. Evaluating the Consequences of Affective Feedback in Intelligent Tutoring Systems. In *Proceedings of 3rd International Conference on Affective Computing & Intelligent Interaction* IEEE Press, 37-42.

ROGOFF, B. 1990. *Apprenticeship in Thinking*. Oxford University Press, New York.

ROMERO, C. AND VENTURA, S. 2007. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*,33, 1, 135-146.

RUMELHART, D., MCCLELLAND, J. AND PDP, R.G. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. MIT Press, Cambridge, MA.

RUS, V. AND GRAESSER, A. 2007. Lexico-syntactic subsumption for textual entailment. In *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, N. NICOLOV, K. BONTCHEVA, G. ANGELOVA AND R. MITKOV, Eds. John Benjamins Publishing Company, Amsterdam, 187-196.

RUS, V., MCCARTHY, P., MCNAMARA, D. AND GRAESSER, A. 2008. A study of textual entailment. *International Journal on Artificial Intelligence Tools*,17, 4, 659-685.

SCHUTZ, P. AND PEKRUN, R. Eds. 2007. *Emotion in Education*. Academic Press, San Diego, CA.

SCHWARTZ, D. AND BRANSFORD, D. 1998. A time for telling. *Cognition and Instruction*,16, 4, 475-522.

SHAH, F., EVENS, M., MICHAEL, J. AND ROVICK, A. 2002. Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions. *Discourse Processes*,33, 1, 23-52.

SLEEMAN, D. AND BROWN, J. Eds. 1982. *Intelligent Tutoring Systems*. Academic Press, New York.

SNOW, R.E., CORNO, L. AND JACKSON, D. 1996. Individual differences in affective and conative functions. In *Handbook of Educational Psychology*, D. BERLINER AND R. CALDEE, Eds. Macmillan, New York, 243–310.

SUSARLA, S., ADCOCK, A., VAN ECK, R., MORENO, K. AND GRAESSER, A. 2003. Development and evaluation of a lesson authoring tool for AutoTutor. In *Supplemental Proceedings of the 11th International Conference on Artificial Intelligence in Education*, V. ALEVEN, U. HOPPE, J. KAY, R. MIZOGUCHI, H. PAIN, F. VERDEJO AND K. YACEF, Eds. University of Sydney School of Information Technologies, Sydney, Australia, 378-387.

SWARTOUT, W.R., GRATCH, J., HILL JR, R.W., HOVY, E., MARSELLA, S., RICKEL, J. AND TRAUM, D. 2006. Toward virtual humans. *AI Magazine*, 27, 2, 96.

VANLEHN, K. 2006. The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*,16, 3, 227-265.

VANLEHN, K., GRAESSER, A., JACKSON, G., 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., JONES, R. AND CHI, M. 1992. A model of the self-explanation effect. *Journal of the Learning Sciences*,2, 1, 1-59.

VANLEHN, K., JORDAN, P., ROSE, C., BHEMBE, D., BOTTNER, M. AND A., G. 2002. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proceedings of the Sixth International Conference on Intelligent Tutoring* S.A. CERRI, G. GOUARDERES AND F. PARAGUACU, Eds. Springer-Verlag, Berlin, 158-167.

VANLEHN, K., LYNCH, C., SCHULZE, K., SHAPIRO, J., SHELBY, R., TAYLOR, L., TREACY, D., WEINSTEIN, A. AND WINTERSGILL, M. 2005. The Andes physics tutoring system: Five years of evaluations. *International Journal of Artificial Intelligence in Education* 15, 147-204.

VANLEHN, K., SILER, S., MURRAY, C., YAMAUCHI, T. AND BAGGETT, W. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction*,21, 3, 209-249.

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

VYGOTSKY, L. 1986. *Thought and Language*. MIT Press, Cambridge, MA.

WAKEFIELD, J.F. 1996. *Educational Psychology: Learning to be a Problem Solver*. Houghton Mifflin, Boston MA.

WANG, N., JOHNSON, W.L., MAYER, R.E., RIZZO, P., SHAW, E. AND COLLINS, H. 2008. The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies*,66, 2, 98-112.

WEINER, B. 1986. *An Attributional Theory of Motivation and Emotion*. Springer-Verlag, New York.

WISHER, R. AND FLETCHER, J. 2004. The case for advanced distributed learning. *Information & Security: An International Journal*,14, 17-25.

WOOLF, B. 2009. *Building Intelligent Interactive Tutors*. Morgan Kaufmann Publishers, Burlington, MA.

WOOLF, B., ARROYO, I., MULDNER, K., BURLESON, W., COOPER, D., DOLAN, R. AND CHRISTOPHERSON, R. 2010. The effect of motivational learning companions on low achieving students and students with disabilities In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, J. KAY AND V. ALEVEN, Eds. Springer, Berlin / Heidelberg, 327-337.

ZENG, Z., PANTIC, M., ROISMAN, G. AND HUANG, T. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,31, 1, 39-58.