Animated pedagogical agents have become very popular in advanced learning environments of the new millennium (Atkinson, 2002; Baylor & Kim, 2005; Graesser, Chipman, Haynes, & Olney, 2005; Hu & Graesser, 2004; Johnson, Rickel, & Lester, 2000; McNamara, Levinstein, & Boonthum, 2004; Moreno, Mayer, Spires, & Lester, 2001; Reeves & Nass, 1996). These agents help students learn by either modelling good pedagogy, by navigating the learner through a complex learning environment, or by holding a conversation in natural language. The agents take on roles of mentors, tutors, peers, players in multiparty games, and avatars in the virtual worlds. They can be designed to have different cognitive abilities, expertise, personalities, physical features, and styles. The agents in some of these systems are carefully scripted and choreographed, whereas agents in other
systems are dynamic and adapt to the learner. The students communicate with the agents through speech, keyboard, gesture, touch panel screen, or conventional input channels. In turn, the agents express themselves with speech, facial expression, gesture, posture, and other embodied actions. In some systems, an ensemble of agents displays social interactions that can model skills of communication. Agents not only enact these strategies, individually or in groups, but can also think aloud while they do so. When an agent reaches the sophistication of having speech recognition and natural language generation, it holds a face-to-face, mixed-initiative dialogue with the student, just as people do in everyday conversations (Cole et al., 2003; Graesser, Chipman, King, McDaniel, & D’Mello, 2007; Gratch et al., 2001; Johnson & Beal, 2005).

This chapter describes AutoTutor, a computer tutor that holds conversations with students in natural language (Graesser, Chipman et al., 2005; Graesser, Lu, et al., 2004; Graesser, Person, & Harter, 2001; Graesser, Oney, Hayes, & Chipman, 2005; Graesser, VanLehn, Rose, Jordan, & Harter, 2001; Graesser, K. Wiemer-Hastings, P. Wiemer-Hastings, Kreuz, & Harter, 1999). AutoTutor simulates the discourse patterns of human tutors and also incorporates a number of ideal tutoring strategies. It presents a series of challenging problems (or questions) that require verbal explanations and reasoning in an answer. It engages in a collaborative, mixed initiative dialog while constructing the answer, a process that typically takes approximately 100 conversational turns. AutoTutor speaks the content of its turns through an animated conversational agent with a speech engine, some facial expressions, and rudimentary gestures. For some topics, there are graphical displays, animations of causal mechanisms, or interactive simulation environments (Graesser, Chipman et al., 2005; Jackson, Olney, Graesser, & Kim, 2006). AutoTutor tracks the cognitive states of the learner by analyzing the content of the dialogue history. AutoTutor dynamically selects the words and statements in each conversational turn in a fashion that is sensitive to what the learner knows. The most current AutoTutor system we are building also adapts to the learner’s emotional states in addition to their cognitive states.

This chapter begins with a description of AutoTutor and the pedagogical principles that motivated its design. We will describe different versions of AutoTutor, including those we are currently working on that are sensitive to the learner’s emotions and that incorporate multiple agents. After describing these systems, we will turn to some important empirical tests of AutoTutor. These include the quality of AutoTutor’s dialogue, learning gains, the students’ perceptions of the learning experience, AutoTutor’s ability to adapt to the learner, and the accuracy of its tracking the cognitive states of the learner.
A SKETCH OF AUTOTUTOR

Figure 5.1 shows a screen shot of AutoTutor on the topic of computer literacy. At the top window is the main question that requires deep reasoning to answer: How is the packet switching model of message transmission like the postal system? An answer to this question involves analogical reasoning, which requires a comparatively deep level of comprehension and application of knowledge (Bloom, 1956; Graesser & Person, 1994). An answer to the question requires approximately one paragraph of information that includes a handful of important statements. It normally takes 50-200 conversational turns between the tutor and student to answer this main question, which is on par with a conversation between a human tutor and a student. The student types in the verbal contribution of each turn by typing in the content, which immediately gets echoed in the window at the bottom of the screen. The conversational agent at the left speaks the content of AutoTutor, with appropriate facial expressions and occasional gestures. The screen shot in Figure 5.1 has an associated diagram for the learner to view. The diagram grounds the content in a visual form that presumably helps learning.

AutoTutor’s dialogues are organized around difficult questions and problems that require reasoning and explanations in the answers. For example,

![Figure 5.1](image-url)  
**Figure 5.1** A computer screen of AutoTutor for the subject matter of introductory computer literacy.
below are two example challenging questions from two of the subject matters that get tutored: Newtonian physics and computer literacy.

PHYSICS QUESTION: If a lightweight car and a massive truck have a head-on collision, upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion, and why?

COMPUTER LITERACY QUESTION: When you turn on the computer, how is the operating system first activated and loaded into RAM?

These questions require the learner to exhibit reasoning in natural language. They are hardly the fill-in-the-blank questions or multiple-choice questions that so many associate with computer-based training. It takes a conversation to answer these challenging questions. For example the conversation below was one of the tutorial interactions with a relatively verbose, knowledgeable student about physics.

TUTOR: Now let’s move on. A lightweight car and a massive truck have a head-on collision. On which vehicle is the impact force greater? Which vehicle undergoes the greater change in motion? Explain why.

STUDENT: The lightweight vehicle experiences the greater impact and the greater change in motion because the mass of the lighter vehicle is less, therefore having less inertia to resist the change in motion. The momentum of the truck is greater than that of the car.

TUTOR: Right! What does Newton’s third law of motion say about the forces that bodies exert on each other?

STUDENT: The forces must be equal in magnitude and opposite in reaction.

TUTOR: Yes, correct! Can you add to that?

STUDENT: The forces must be equal and opposite reactions acting on opposite bodies.

TUTOR: Ya! We’ve almost covered it, but there’s one part I think is still missing. I am sure you can get this. During an interaction, each body exerts an equal but opposite force on the other __________?

STUDENT: body

TUTOR: Yes, right! on the other body. So when two balls, one plastic and the other iron, hit each other, does the heavier or neither ball exert the greater impact force on the other?

STUDENT: neither
TUTOR: Yes, correct! Neither, the forces are equal in magnitude. Let’s take a moment for general physics questions. Ask me one and I’ll do my best to answer it.

STUDENT: no

TUTOR: Alright. To recap what we have just discussed, please give me your complete answer to the original question, as presented at the top of the screen.

STUDENT: Although the lighter car experiences a greater change in motion, both vehicles exert an equal impact force, but opposite in direction. The larger object has a larger mass, and therefore a larger inertial resistance than that of the car.

TUTOR: That is right! Putting everything together, um, According to Newton’s third law of motion, if one body exerts a force on the other body then the other body must exert an equal and opposite force on the first body. Therefore, the car and the truck both experience the same impact force. By Newton’s second law, the car will experience the greater change in its motion because its mass is much smaller. This question is finished so let’s go on to the next question.

THE DIALOGUE MOVES OF AUTOTUTOR

The example dialogue above conveys the flavor of what a conversation would be like with AutoTutor, at least for a verbose and knowledge student. However, many students are not highly verbal and knowledgeable about physics. In fact, when students are asked these types of questions, their initial answers are typically only one or two sentences in length. However, one to two sentences provide insufficient information to adequately answer the question so tutorial dialogue is needed to flesh out a complete answer. AutoTutor engages the student in a mixed-initiative dialogue that draws out more of what the student knows and that assists the student in the construction of an improved answer. So how does AutoTutor converse with the learner? There are a number of dialogue moves that systematically steer the interaction.

Feedback

AutoTutor provides feedback to the student on the quality of what the student contributes in their previous turn before AutoTutor responds. The feedback may be positive (very good, bravo), negative (not quite, almost), neutral (uh huh, okay), or exhibit even finer gradations. Sometimes the verbal
expressions are neutral, but the facial expressions are skeptical, a pattern of expression that human tutors often exhibit (Fox, 1993; Graesser & Person, 1994; Person, Kreuz, Zwaan, & Graesser, 1995). In essence, the tutor wants to be verbally supportive so that the student does not shut down from negative feedback, whereas subtle facial expressions or pauses give the accurate pedagogical feedback. It is possible to give AutoTutor an interesting personality by systematically selecting the feedback expressions. For example, we have developed a rude tutor that provides feedback with an edge, that is, *I sure pegged you wrong after that pathetic answer* (negative feedback) or *Aren’t you the little genius* (positive feedback). Most adults have more fun interacting with the rude tutor than the original earnest prude tutor.

Covering the Answer

AutoTutor has dialogue moves that advance the goal of answering the main question. These dialogue moves include *pumps* for the student to provide information (e.g., What else?), *hints* for the student to generate correct ideas, *prompts* for the student to fill in missing words, *assertions* that fill in ideas that the student leaves out, and *summaries* of the complete answer near the end of the exchange. These acts of pumps, hints, prompts, assertions, and summaries vary along a continuum from getting the student to supply information to the tutor supplying information. According to constructivist theories of tutoring and pedagogy, it is best to get the student to do the talking and actively construct the explanatory reasoning rather than the tutor serving as a mere information delivery system (Aleven & Koedinger, 2002; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; McNamara, 2004). AutoTutor starts out the exchange giving pumps and hints, which tend to be sufficient for the more knowledgeable, verbose students (Jackson & Graesser, 2006). AutoTutor resorts to prompts and assertions when the student cannot make progress in constructing the correct information.

Corrections of Errors and Misconceptions

When AutoTutor identifies an erroneous idea or misconception expressed by the student, it generates a *correction*. Most versions of AutoTutor immediately correct the erroneous information, following the design of intelligent tutoring systems (Anderson, Corbett, Koedinger, & Pelletier, 1995; VanLehn et al., 2002). The alternative would be an indirect Socratic tutor that asks the students well-selected questions that lead the students to self-discover their own misconceptions. However, such Socratic questions require a level of intelligence and subtlety that human tutors rarely exhibit.
(Graesser, Magliano, & Person, 1995). There are some versions of AutoTutor, however, that launch interactive simulations (Jackson et al., 2006) or embedded mini-dialogues (VanLehn et al., 2007) when misconceptions are detected.

**Answers to Student Questions**

Students do not frequently ask questions in classrooms whereas the rate increases in tutoring environments (Graesser, McNamara, & VanLehn, 2005). For example, the number of student questions is only one question per six to seven hours in a classroom, but increases to one question per two minutes in human tutoring (Graesser & Person, 1994). AutoTutor attempts to answer student questions when they occur during the tutoring session. Definitional questions (*What does X mean?*) are among the most frequent question categories that students ask. AutoTutor answers these questions by matching the queried word to entries in an electronic dictionary and producing the answer. For most other question categories, AutoTutor searches for a highly relevant paragraph in an electronic textbook for an answer. Sometimes AutoTutor does not understand the student’s question and asks the student to rephrase the question (e.g., *I don’t quite understand the question, so could you ask it with different words?*). AutoTutor also announces when it cannot answer a question (*That’s a good question but I cannot answer it now*) and inquires whether the student finds the answer satisfactory (*Does this answer your question?*). Although AutoTutor has these question answering facilities, students in the AutoTutor sessions still ask a disappointing number of questions. Graesser, McNamara, and VanLehn (2005) reported that only 13.2 student questions are asked per 100 turns. This is about twice the number of student questions when college students interact with expert human tutors on the same topic. Even college students have inadequate skills of self-regulated learning (Azevedo & Cromley, 2004) so they rarely ask questions to fill in their knowledge gaps. As a consequence, both human tutoring and AutoTutor sessions are not organized around student questions and inquiry. It is the tutor who manages most of the tutorial agenda. For this reason, AutoTutor periodically invites the student to ask questions, as it does in the example dialogue presented earlier.

**THE STRUCTURE OF TUTORIAL DIALOGUE**

The structure of the dialogue in both AutoTutor and human tutoring (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Graesser et al., 1995; Shah, Evens, Michael, & Rovick, 2002) can be segregated into three levels or as-
pects: (1) expectation and misconception-tailored dialogue, (2) a 5-step dialogue frame, and (3) composition of a conversational turn. These three levels can be automated and produce respectable tutorial dialogue.

**Expectation and Misconception Tailored Dialogue**

This is the primary pedagogical method of scaffolding good student answers. Both AutoTutor (Graesser et al., 2005) and human tutors (Graesser et al., 1995) typically have a list of anticipated *expectations* (good answers) and a list of anticipated *misconceptions* associated with each challenging question or problem in the *curriculum script* for a subject matter. For example, listed below are the expectations (E1, E2, E3) and misconceptions (M1 through M5) that are relevant to the example physics problems.

**Expectations**

(E1) The magnitudes of the forces exerted by the two objects on each other are equal.
(E2) If one object exerts a force on a second object, then the second object exerts a force on the first object in the opposite direction.
(E3) The same force will produce a larger acceleration in a less massive object than a more massive object.

**Misconceptions**

(M1) A lighter/smaller object exerts no force on a heavier/larger object.
(M2) A lighter/smaller object exerts less force on other objects than a heavier/larger object.
(M3) The force acting on a body is dependent on the mass of the body.
(M4) Heavier objects accelerate faster for the same force than lighter objects.
(M5) Action and reaction forces do not have the same magnitude.

AutoTutor guides the course of covering the expectations through the dialogue moves described earlier (pumps, hints, prompts, assertions, summaries). Hints and prompts are carefully selected from the curriculum script by AutoTutor to elicit content that fills in missing content words, phrases, and propositions. For example, a hint to get the student to articulate expectation E1 might be “What about the forces exerted by the vehicles on each other?”; this hint would ideally elicit the answer “The magnitudes of the forces are equal.” A prompt to get the student to say “equal” would be “What are the magnitudes of the forces of the two vehicles on each other?” As the learner expresses information over many turns, the list of expectations is eventually
covered and the main question is scored as answered. Complete coverage of the answer therefore requires AutoTutor to have a pool of hints and prompts available in the curriculum script to extract all of the content words, phrases, and propositions in each expectation. AutoTutor adaptively selects those hints and prompts that fill missing constituents and thereby achieves pattern completion. We have also developed question generation mechanisms that automatically generate hints and prompts instead of requiring lesson planners to prepare these pools by hand (Cai et al., 2006).

AutoTutor is dynamically adaptive to students in ways other than coaching them to articulate expectations. There are conversational goals of correcting misconceptions that arise in the student’s talk, of giving feedback on the quality of the student’s previous turn, and of answering the student’s questions, as discussed earlier.

**Five-Step Dialogue Frame**

This dialogue frame is prevalent in human tutoring (Graesser & Person, 1994; VanLehn et al., 2007) and is implemented in AutoTutor. The five steps of the dialogue frame are:

1. Tutor asks the main question.
2. Student gives an initial answer.
3. Tutor gives short feedback on the quality of the student’s answer in #2.
4. Tutor and student collaboratively interact via expectation and misconception tailored dialogue.
5. Tutor verifies that the student understands (e.g., *Do you understand?*)

Students often respond that they understand the answer in step five, when in fact many do not understand because they have inadequate metacognitive knowledge (Graesser & Person, 1994; Maki, 1998). A good tutor would probe the student further by asking more penetrating questions to diagnose the student’s understanding, but even good tutors rarely do this. Most tutors end up giving a summary answer to the main question and then select another main question. A good tutor would request the student to provide the summary (as in our example dialogue) rather than it being provided by the tutor, but even good tutors rarely do that.

**Managing One Conversational Turn**

Each turn of AutoTutor in the conversational dialogue has three information slots (i.e., units, constituents). The first slot of most turns is short
feedback on the quality of the student’s last turn. The second slot advances the coverage of the expectations with either pumps, hints, prompts, assertions, corrections, answers to student questions, requests for summaries, or presentation of summaries. The third slot is a cue to the student for the floor to shift from AutoTutor as the speaker to the student. For example, AutoTutor ends each turn with a question or a gesture to cue the learner to do the talking. Multiple dialogue moves occur in most of AutoTutor’s turns. Therefore, discourse markers (and also, okay, well, perhaps you can answer this question) connect the utterances of these three slots of information within each turn.

The three structural levels of AutoTutor go a long way in simulating a human tutor. AutoTutor can keep the dialogue on track because it is always comparing what the student says to anticipated input (i.e., the expectations and misconceptions in the curriculum script). Pattern matching operations and pattern completion mechanisms drive the comparison. These matching and completion operations are based on latent semantic analysis (Landauer, McNamara, Simon, & Kintsch, 2007) and symbolic interpretation algorithms (Rus & Graesser, 2006) that are beyond the scope of this article to address. AutoTutor cannot interpret student contributions that have no matches to anticipated content in the curriculum script (i.e., the expectations and misconceptions associated with the main question). This of course limits true mixed-initiative dialogue. That is, AutoTutor cannot explore the topic changes and tangents of students as the students introduce them. However, available studies of naturalistic tutoring (Chi et al., 2001; Chi, Siler, Jeong, 2004; Graesser et al., 1995) reveal that (a) human tutors rarely tolerate true mixed-initiative dialogue with student topic changes that steer the conversation off course and (b) most students rarely change topics, ask questions, and spontaneously grab the conversational floor. Instead, it is the tutor that drives the dialogue and leads the dance. AutoTutor and human tutors are very similar in these respects.

VERSIONS OF AUTOTUTOR

Different versions of AutoTutor have been designed to incorporate particular pedagogical goals and cover different subject matters. The subject matters covered so far are computer literacy, physics, biology, tactical planning, and critical thinking. In the version on introductory computer literacy, for example, AutoTutor covered hardware, operating systems, and the internet. Each of these topics had 12 challenging questions that required deep reasoning, such as why, how, what-if, what if not, how is X similar to Y?. In most versions of AutoTutor, the students type in their contributions via keyboard, whereas recent versions allow spoken input (Graesser et al.,
2007). We have used the commercially available Dragon Naturally Speaking™ (version six) speech recognition system for speech-to-text translation. The interface on some versions of AutoTutor includes a dialogue window that presents the history of the turn-by-turn tutorial dialogue for the student to inspect during tutoring. The students can scroll back as far as they want in this dialogue history, although very few students pursue this dialogue recovery.

Most versions of AutoTutor have animated conversational agents 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, Moreno 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 et al., 2003). Simply put, the medium is not the message, but rather the message is the message.

**AutoTutor-3D**

One version of AutoTutor, called *AutoTutor-3D*, guides learners on using interactive simulations of physics microworlds (Graesser, Chipman et al., 2005; Jackson et al., 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 5.2 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.

**Aries**

One recent version of AutoTutor, called *ARIES* (Acquiring Research Investigative and Evaluative Skills) is currently being developed in a collaboration between University of Memphis, Northern Illinois University, and Claremont
ARIES teaches scientific critical thinking with two animated pedagogical agents. The potential of agents taking on different social or pedagogical roles has been investigated by previous researchers (Baylor & Kim, 2005; P.Wiemer-Hastings & Graesser, 2000). 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 indeed been pursued in the Teachable Agent research of Biswas, Leelawong, Schwartz, and Vye (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 (perhaps called a triologue) among the human student, the expert agent, and the student agent. We have already completed an initial experiment that shows some promise in ARIES facilitating learning compared to normal training methods for scientific critical thinking.
An Emotion-Sensitive AutoTutor

We are currently working on a version of AutoTutor that is sensitive to the student’s emotions. AutoTutor is augmented with sensing devices and signal processing algorithms that classify affective states of learners (see Figure 5.3). Emotions are classified on the basis of dialog patterns during tutoring, the content covered, facial expressions, body posture, and speech intonation (D’Mello, Craig, & Graesser, 2006; D’Mello, Picard, & Graesser, in press; Graesser, Jackson, & McDaniel, 2007). Our previous research has revealed that the primary emotions that occur during learning with AutoTutor are frustration, confusion, boredom, and flow (intense engagement), whereas surprise and delight occasionally occur (Craig, Graesser, Sullins, & Gholson, 2004; Graesser, Witherspoon et al., 2006). The accuracy of the computer detecting and classifying these emotions on the basis of dialogue history is not perfect (approximately .33 to .50 on a metric that varies from 0 to 1 reliability) and depends somewhat on the emotion (D’Mello, Craig, & Graesser, 2006; D’Mello, Craig, Witherspoon, McDaniel, & Graesser, in press). These percentages are expected to increase with the addition of the other channels of communication, namely facial expressions (Ekman, 2003; McDaniel et al., 2007), body posture (D’Mello, Chipman, & Graesser, 2007; Kapoor & Picard, 2003), and speech intonation (Litman & Forbes-Riley, 2004). D’Mello, Picard, and Graesser (in press) have reported the accuracy of detecting and classifying the major emotions during learning with AutoTutor on the basis of dialogue history, facial expressions, body posture, and different combinations of these channels. We are confident that these detection and classification accuracies are sufficiently accurate that it would be feasible to build an affect-sensitive AutoTutor.

The next step in this research is to explore whether learning gains and learner’s impressions of AutoTutor are influenced by dialogue moves of AutoTutor that are sensitive to the learner’s emotions. For example, if the student is extremely frustrated, then AutoTutor presumably should give a good hint or prompt that directs the student in a more positive learning trajectory. If the student is bored, AutoTutor should give more engaging, challenging, and motivating problems. If the student is very absorbed and satisfied, then AutoTutor should be minimally directive. The emotion exhibited by AutoTutor is also an important consideration, just as it is for human tutoring (Lepper & Henderlong, 2000). Should AutoTutor be empathetic to a frustrated student, or be earnest, forceful, or upbeat? Answers to such questions await future research.
Figure 5.3. Communication channels for detecting emotions of the learner.
Authoring New Content for AutoTutor

AutoTutor was designed to be generic, rather than domain-specific. AutoTutor is appropriate for any subject matter that involves verbal reasoning, as opposed to precise, closed-world knowledge and skills such as mathematics, statistics, or logic. There is an authoring tool that makes it easy for instructors and curriculum designers to prepare new material on new topics in a curriculum script. It is called the AutoTutor Script Authoring Tool, ASAT (Susarla, Adcock, Van Eck, Moreno, & Graesser, 2003). ASAT guides the curriculum designer through the process of creating content. The person does not need to have any advanced knowledge of computer software in order to use this authoring tool. We should acknowledge, however, that it has always been extremely difficult to develop authoring tools for advanced learning environments (Murray, Blessing, & Ainsworth, 2003). It would be misleading to say that the curriculum designer would not benefit tremendously from knowledge of information technologies, education, language, and cognitive science.

A basic version of AutoTutor can be developed for a new topic in three steps. First, a latent semantic analysis (LSA) space is constructed from a large corpus of electronic texts on the subject matter. Actually, this step may be circumvented by simply using an existing extremely large corpus and LSA space that covers a broad landscape of subject matters (see Landauer et al., 2007). Second, a glossary of terms is integrated with AutoTutor. Electronic glossaries are routinely available for most subject matters. Third, the lesson planner creates the curriculum script with ASAT, which methodically guides the person through the process of creating content.

The curriculum script consists of a questions or problems that would stimulate a dialogue between AutoTutor and the student. For example, the structure of the curriculum script for our physics tutor had 10 problems that we embedded in the material the students read on physics. The structural components of the curriculum script are presented below for the physics tutor version that did not have embedded interactive simulations.

1. Set of physics problems (topics)
2. Problem → Context + Question + Explain + (Picture)
3. Ideal answer with full line of reasoning in paragraph
4. Set of important expectations in ideal answer
   a. A family of hints, prompts, and assertions to elicit expectations
   b. A hint completion for each hint (i.e., correct response to a hint)
   c. A prompt completion for each prompt
5. Set of anticipated misconceptions
   a. Plus a correction for each misconception
6. Set of anticipated questions and their answers
7. Set of key concepts
   a. Plus one more functionally equivalent referring expressions for each concept
8. Summary of ideal answer
9. Mark-up language for animated agents, with pauses, durations of words, emphasis, gestures, emotions, and so forth

An expanded ASAT would be needed for external media with pictures to point to and interactive simulations to manipulate. As with all authoring tools, ASAT has instructions, help, examples, and other forms of scaffolding to assist the curriculum designer.

At this point ASAT has been used by approximately 30 students, faculty, and teachers. We completed a systematic study on 15 experts in biology who created curriculum scripts on a sample of topics. It took approximately 40 minutes for a lesson planner to create the curriculum script for one difficult problem or question that takes a typical student 20 minutes to answer. Developing AutoTutor for a new course could be accomplished in a time span measured in weeks or months. This is much faster than many intelligent tutoring systems that take years to develop.

**EMPIRICAL EVALUATIONS OF AUTOTUTOR**

We have tested AutoTutor on a number of empirical criteria. These include the quality of AutoTutor’s dialogue, learning gains, the learner’s perceptions of the learning experience with AutoTutor, AutoTutor’s ability to adapt to the learner, and the accuracy of AutoTutor’s tracking of the cognitive states of the learner. The section gives the highlights of what we have learned from these empirical evaluations of AutoTutor.

**Quality of AutoTutor’s Dialogue**

The conversations managed by AutoTutor are hardly perfect, but are smooth enough for students to get through the sessions with minimal difficulties. Indeed, we have been consistently surprised that no student, among the thousands of students we have run, has left the tutoring session with frustration, irritation, or disgust. The quality of the conversation will only continue to improve as we add deeper natural language understanding facilities to the system that reflect advances in computational linguistics (Jurafsky & Martin, 2000), discourse processes (Graesser, Gernsbacher, & Goldman, 2003), and the representation of world knowledge (Landauer et al., 2007).
One series of studies on dialogue quality performed a fine-grained bystander Turing test. These studies revealed that AutoTutor’s dialogue is sufficiently tuned so that a bystander who observes tutorial dialogue in print cannot tell whether a particular turn was generated by AutoTutor or by an expert human tutor of computer literacy (Person & Graesser, 2002). How was this type of Turing test conducted? A series of studies were conducted that randomly sampled AutoTutor’s turns within a tutorial dialogue session with AutoTutor. Half of the turns were generated by AutoTutor and half were substituted by a human expert tutor on the basis of the dialogue history. Bystander participants were presented these tutoring moves in a written transcript and asked to decide whether each sampled turn was generated by a computer or a human. The results revealed that the bystander judges could not tell the difference. The judges claimed that human-generated tutor moves were generated by humans 49% of the time and AutoTutor-generated moves were generated by humans 50% of the time. AutoTutor therefore successfully passed the fine-grained bystander Turing test for individual tutoring turns. It should be emphasized, however, that a bystander can eventually tell whether a sequence of turns was part of a dialogue with AutoTutor versus a human tutor. AutoTutor’s dialogue is hardly perfect because AutoTutor does not have the depth of language comprehension and global coherence. But surprisingly, AutoTutor is close enough to human tutorial dialogue to keep the conversation going.

We have conducted very detailed analyses of AutoTutor that point to aspects of the dialogue and tutoring mechanism that could be improved. One problem lies in errors in interpreting the content of student turns. The pattern matching operations between student contributions in a turn and AutoTutor’s expectation statements (i.e., E1, E2, E3) are not perfect, so some students get frustrated and conclude that AutoTutor is not listening. We are currently improving the language analyses facilities at the level of syntax, semantics, and inference generation (Rus & Graesser, 2006).

A second problem consists of misclassification of the speech acts in student turns. The student turns are segmented into speech acts and each speech act is assigned to one of approximately 20 speech act categories. These categories include student assertions, questions in 16 different categories, short responses (yeah, right), meta-cognitive expressions (I don’t understand, I see), and meta-communicative expressions (What did you say?). The accuracy of classifying the student speech acts into categories varies from .87–.96 (Olney et al., 2003), but is not perfect. The dialogue coherence breaks down when some misclassification errors occur, which ends up confusing students. More efforts are needed to improve the speech act classification accuracy and to manage the dialogue to minimize exposure of unwanted consequences.
A third problem occurs when the AutoTutor does not generate relevant and informative answers to the student questions. AutoTutor can handle roughly half of the student questions so half of AutoTutor’s replies are either incorrect, constitute requests for clarification (I don’t understand your question, so could you rephrase it?), or pass the burden onto the student (That’s a good question, so how would you answer it?). Improvements in the question answering facilities are needed to minimize this third problem.

A skeptic may raise the objection that animated pedagogical agents like AutoTutor will ultimately be a disaster when they do not completely understand the human (Shneiderman & Plaisant, 2005), but create expectations in the human that they do understand (Norman, 1994). This is true when there are high expectations on the quantitative precision of answers and there is a high degree of shared knowledge between humans and computers. However, this is not the case when tutoring on verbal content for students who have little or no subject matter knowledge and when the tutor hedges on how well it understands the student. An AutoTutor might be just the right fit for this niche.

We have recently tested the first version of AutoTutor with speech recognition. The Dragon speech recognition system is correctly translating roughly 75% of the content words so AutoTutor’s semantic evaluator was far from perfect. Nevertheless, AutoTutor produced reasonable responses to most of what the student said, even though its understanding was imperfect. We suspected that many of the college students who used this version of AutoTutor had the illusion that AutoTutor was comprehending them. But indeed, that just may be the way it is when students try to communicate with human tutors. The real test of the value of AutoTutor does not lie in its ability to comprehend perfectly, but rather in its comparisons to humans and in its ability to facilitate learning.

Learning Gains with AutoTutor

One might ask the fundamental question of why we would expect learning gains to be comparatively high in a conversation-based tutor like AutoTutor? Why would we expect AutoTutor to be better than reading a textbook, classroom teaching, or a traditional computer-based training system that is adaptive and that gives immediate feedback?

There are many reasons to expect AutoTutor to show better learning gains than reading a textbook and classroom instruction. For example, AutoTutor’s emphasis on deep reasoning and challenging problems to solve rather than acquiring shallow knowledge and facts is one obvious reason. These deeper levels of processing are not typically achieved in normal reading and classroom experiences. Another reason is that conversation
in natural language is more accessible to many learners than other forms of media and communication. A significant segment of the population of learners prefer face-to-face communication in natural language over texts, mathematical formulae, graphs, tables, and other types of media. A different angle to answering the above questions would be to acknowledge the empirical research on human tutoring that has shown its advantages over classroom learning (Cohen, Kulik & Kulik, 1982) and reading textbooks (Graesser et al., 2004; VanLehn et al., 2007). AutoTutor attempts to mimic the discourse patterns of human tutors so it should be expected to enjoy such advantages in learning gains.

At this point in time, there have been no head-to-head comparisons between AutoTutor and conventional computer-based training systems that deliver adaptive instruction and immediate feedback. There have also been no comparisons with sophisticated intelligent tutoring systems that lack natural language communication. Research is needed to determine how the advantages of natural language dialogue and the pedagogical mechanisms of AutoTutor compare to the features of other systems. AutoTutor does not handle precise mathematical computations so intelligent tutoring systems might show an advantage on those dimensions even though they pay a penalty by lacking natural language dialogue. It is plausible that there are aptitude-treatment interactions, with (a) lower ability learners showing greater benefits from AutoTutor’s conversational component but (b) high ability students showing greater benefits from a conventional computer-based system with fast efficient feedback or intelligent systems that can adaptively reason with mathematical precision.

The learning gains of AutoTutor have been evaluated in roughly 20 experiments conducted during the last decade. Assessments of AutoTutor on learning gains have shown effect sizes of approximately .8 standard deviation units in the areas of computer literacy (Graesser et al., 2004) and Newtonian physics (VanLehn, Graesser et al., 2007) when contrasted with appropriate comparison conditions. An effect size with the value of 1 standard deviation (called a sigma) is approximately a one letter grade difference. These evaluations of AutoTutor place AutoTutor somewhere between untrained human tutors who yield effect sizes of .42 sigma (Cohen, Kulik, & Kulik, 1982) and an intelligent tutoring system with ideal tutoring strategies that yield effect sizes of 1.0 sigma (Corbett, 2001). The assessments of learning gains from AutoTutor have varied between 0 and 2.1 sigma (a mean of .8), depending on the learning performance measure, the comparison condition, the subject matter, and the version of AutoTutor.

Approximately a dozen measures of learning have been collected in these assessments of AutoTutor. The measures include: (1) multiple choice questions on shallow knowledge that tap definitions, facts and properties of concepts, (2) multiple choice questions on deep knowledge that taps
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 subjects fill in missing words of texts that articulate explanatory reasoning on the subject matter, and (5) performance on problems that require problem solving. Attempts are also made to control for training time among the various conditions in these assessments of learning gain.

Assessments of learning in these various conditions have uncovered a number of findings that are summarized in previous published studies (Graesser, Lu et al., 2004; VanLehn et al., 2007). Some of these findings are counterintuitive.

1. **AutoTutor versus reading a textbook.** Learning gains with AutoTutor are superior to reading from a textbook on the same topics for an equivalent amount of time. This result applies to tests of deep knowledge (i.e., explanations and inferences, as in the Force Concept Inventory, Hestenes, Wells, & Swackhamer, 1992) rather than shallow knowledge. One might expect a textbook to yield superior learning because the learner has absolute control over inspection strategies. However, the flip side is that most readers settle for shallow standards of comprehension (Hacker et al., 1998; Maki, 1998) so they do not engage in the acquisition of deep knowledge from reading.

2. **Reading a textbook versus doing nothing.** Learning gains are zero in both of these conditions when the tests tap deeper levels of comprehension. This may be because there is a very low correlation \( r = .27 \) between college students’ perceptions of how well they are comprehending and their actual comprehension when measured by objective tests (Maki, 1998). Readers need difficult problems, such as those posed by AutoTutor, that challenge their *illusions of comprehension* (Glenberg, Wilkinson, & Epstein, 1982); this does not occur when students read text with shallow standards of comprehension.

3. **AutoTutor versus expert human tutors.** Learning gains from AutoTutor have been compared with the gains of accomplished human tutors who communicated with the students through computers in computer-mediate communication, as opposed to face-to-face. Learning gains were equivalent for students with a moderate degree of physics knowledge. In contrast the expert human tutors prevailed when the students had low physics knowledge and the dialogue was spoken.

4. **Deep versus shallow tests of knowledge.** The largest learning gains from AutoTutor have been on deep reasoning measures rather than measures of shallow knowledge (e.g., definitions of terms, lists of entities, properties of entities, recognition of explicit content).
5. **Zone of proximate development.** AutoTutor is most effective when there is an intermediate gap between the learner’s prior knowledge and the ideal answers of AutoTutor. AutoTutor is not particularly effective in facilitating learning in students with high domain knowledge and when the material is too much over the learner’s head.

One way of analyzing the learning gains is to compare the normal conversational AutoTutor with different comparison conditions. We computed mean effect sizes for these contrasts on multiple choice questions that tapped deep reasoning. The conversational AutoTutor has (a) a .80 effect size (sigma) compared with pretests, reading a textbook, or doing nothing, (b) a .22 sigma compared with reading text book segments directly relevant to the AutoTutor problems, (c) a .13 sigma compared with AutoTutor presenting speech acts in print instead of the talking head, (d) a .08 sigma compared with expert human tutors in computer-mediated conversation, and (e) a –.20 sigma compared with a version of AutoTutor that is enhanced with interactive 3D simulations (i.e., the interactive simulations are better). Regarding AutoTutor 3D, the interactive simulation is helpful only for the learners who actually use the interactive microworld (Jackson & Graesser, 2006).

**The Learner’s Perception of the Learning Experience**

Jackson and Graesser (2007) recently conducted a study with different versions of AutoTutor that assessed both learning gains and the learner’s perceptions of the system. The learning measures consisted of multiple choice questions that tapped deep comprehension and explanation-based reasoning about Newtonian physics. The perception measures were rating scales on a number of dimensions: their liking of the learning experience, the ease of interacting with the system, how much they believed they learned, and how interesting the experience was. Each of these ratings was on a 6-point scale.

Jackson and Graesser (2007) had different versions of AutoTutor that manipulated the feedback that the college students received during their interactions with AutoTutor. In one condition, they received both content and progress feedback. The content feedback consisted of (a) highlighting in red the important words about physics (e.g., force, acceleration, mass) in the dialogue history after the students expressed them in their conversational turns and (b) presenting a summary of the ideal answer to the question after the interaction was finished for that question. The progress feedback consisted of (a) progress bars on how much each expectation was covered after each student turn and (b) points on their performance after each question was answered. In ad-
dition to this Content+Progress Feedback condition, students were randomly assigned to a Content Feedback only condition, a Progress Feedback only condition, and a No Feedback condition. Our prediction was that the feedback would improve the student’s perceptions of the learning experience.

The results of the experiment were quite illuminating in a number of respects. We were happy to confirm that learning gains were significantly improved by the feedback when we analyzed proportional learning gains, \[ \frac{(\text{posttest score} - \text{pretest score})}{(1 - \text{pretest score})} \]. These scores were 0.19, 0.28, 0.14, and 0.07 in the Content+Progress, Content, Progress, and No Feedback conditions, respectively. 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. For example, the Content Feedback condition yielded much higher learning than the No Feedback condition, but the students’ mean interest ratings were substantially lower in the Content Feedback condition than the No Feedback condition, 3.25 versus 4.73, respectively. 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. Such a result of course calls into question the role of student perceptions in the design of learning environments that attempt to maximize learning.

When considering all of the studies we have conducted on students’ perceptions of AutoTutor, the best conclusion to be made is that the ratings are average, but slightly lean in the positive direction. College students are not gaga over the talking head and conversational facilities that we have designed. Perhaps their impressions would improve with a better talking head and an improved engineering of the dialogue. According to the available empirical studies in our lab, student impressions are significantly affected by the quality of the speech engine (Louwerse, Graesser, Lu, & Mitchell, 2005) and particular properties of the facial persona that are not well understood (Graesser, Moreno et al., 2003; Moreno, Klettke, Nibbaragandla, Graesser, 2002), but it is the quality of the content in the conversational turns that matters most (Graesser, Moreno, et al., 2003). Therefore, we are currently investing most of our effort in improving the computational linguistics components on natural language understanding (Rus & Graesser, 2006; Rus, McCarthy, & Graesser, 2006) and natural language generation (Cai et al., 2006; Rus, Cai, & Graesser, 2007).

**AutoTutor’s Adaptation to the Student**

AutoTutor was designed to be adaptive to the subject matter knowledge of the learner, the verbal abilities of the learner, and the dialogue history.
More specifically, the dialogue moves composed by AutoTutor are sensitive to the following constraints and parameters: The speech act categories of the student in the previous turn (e.g., are they asking a question or asserting information), the quality of the student’s assertions in the previous turn, the coverage of the expectations throughout the dialogue history for the particular problem under focus, the words that the student has articulated in the expectation under focus, misconceptions expressed by the student, and a dialogue planning module. Therefore, no two conversations with AutoTutor are ever the same. AutoTutor is an adaptive, dynamic system; it is not a rigid, scripted, choreographed, information-delivery system.

Jackson and Graesser (2006) performed analyses that assessed the sensitivity of AutoTutor to the subject matter knowledge of the learner. Pretest scores on the subject matter (physics or computer literacy) were available as a measure of how much prior knowledge the student has on the domain being tutored. If AutoTutor is truly adaptive, then the pretest scores should predict how often AutoTutor gives particular types of feedback, corrections, and dialogue moves that attempt to cover the expectations (namely the pumps, hints, prompts, and assertions). Correlational analyses have confirmed the obvious predictions that would be made if AutoTutor were sensitive to the knowledge of the learner. Consider first the short feedback that AutoTutor gives to the student after most of the student’s turns; this short feedback is either positive (very good), neutral (okay), or negative (not quite). As would be expected, the pretest scores on student domain knowledge had a significant positive correlation with AutoTutor’s positive short feedback moves and a negative correlation with negative feedback. Next consider the corrections that AutoTutor makes when identifying student errors/ misconceptions and subsequently correcting them. There was a negative correlation between pretest scores and the frequency of corrections by AutoTutor. Finally, consider 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 student knowledge that reflects a continuum from the student supplying information to the tutor supplying information as we move from pumps, to hints, to prompts, to assertions. The correlations were indeed positive between pretest scores and the proportion of dialogue moves that are pumps and hints, but negative with prompts and assertions. For students with more subject matter knowledge, 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 subject matter knowledge, AutoTutor needs to generate prompts for the student to articulate specific words, or alternatively 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 modicum of accuracy and adaptively responds to the learner’s level of knowledge.

**Tracking the Cognitive States of the Learner**

We have conducted a series of studies that assess how well AutoTutor tracks the particular expectations associated with a good answer to a challenging question in physics or computer literacy (Graesser, Penumatsa, Ventura, Cai, & Hu, 2007; Rus & Graesser, 2006). As discussed earlier, AutoTutor is continuously monitoring whether the student’s contributions match three to seven expectations (e.g., E1 in the example presented earlier is “The magnitudes of the forces exerted by the two objects on each other are equal”) and three to seven misconceptions (e.g., M1 is “A lighter/smaller object exerts no force on a heavier/larger object”). A *match evaluation score* for each expectation and misconception is continuously being updated after each student turn. The match evaluation score varies from 0 to 1, with 1 signifying that all of the semantic content in the expectation/misconception is covered by the contributions expressed by the student during the tutorial dialogue. The computational algorithms of the match evaluation scores have varied among versions of AutoTutor and will not be addressed in this chapter. It suffices to say that these algorithms consider inferential knowledge in addition to coverage of explicit content words in the expectation/misconception.

We have analyzed the accuracy of the match evaluation scores by comparing AutoTutor’s scores to judgments of subject matter experts. For example, we have analyzed the complete answers that students give as an answer to one of the challenging physics questions, recorded AutoTutor’s match evaluation score for each expectation/misconception, and collected ratings from five expert physicists as to whether each expectation/misconception was present in the student answers. There were three values in these expert judgments: (1) explicitly present, (2) implicitly present, versus (3) absent. A stringent criterion was a rating of 1 whereas a lenient criterion was a rating of one or two. Similarly, misconceptions were judged on a 3-point scale. When considering all judges, an expectation or misconception was considered covered if the majority of experts scored it as covered (given the operative criterion). The correlations between these match evaluation scores and expert ratings have varied between .35 and .50, depending on the criterion, 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 five ran-
randomly sampled student answers per expectation, yielding a total of 125 pairs of expressions. This analysis involves comparisons between sentential expectations and single student speech acts during training. The units of comparison are not the same as previous analyses in which an expectation is compared with content of an entire student answer. 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. We scaled the 125 items on two other metrics to see how they would compare with AutoTutor’s original match evaluation score, which was based entirely on latent semantic analysis (Landauer et al., 2007). First, we computed overlap scores between the content words in the student speech acts and expectations. If an expectation has \(A\) content words, a student speech act has \(B\) content words, and there are \(C\) common words between the two sentential units, then the overlap score is computed as \(\frac{2C}{(A+B)}\). The correlation between expert ratings and word overlap scores was .39, so a simple word overlap metric does a reasonable job when sentential units are compared. Second, we scaled whether the serial order of common content words was similar between the two sentential units by computing Kendall’s Tau scores; this word order similarity metric had a .25 correlation with the expert ratings. We performed a multiple regression analysis that assessed whether the expert ratings could be predicted by LSA, word overlap, and Kendall’s Tau together. The three predictors accounted for a significant \(r = .42\). Therefore, we believe that analyses of sentences would benefit from a computational algorithm that considers a mixture of several interpretation mechanisms, including semantic entailment and inference (Rus, McCarthy, & Graesser, 2006). This will be an important priority in future research because some students get frustrated when they feel that AutoTutor does not understand them.

**Tracking the Emotional States of the Learner**

A series of studies have assessed how well AutoTutor classifies whether the learner is experiencing frustration, confusion, boredom, flow/engagement, delight or surprise (D’Mello et al., 2006; D’Mello, Picard, & Graesser, 2007; D’Mello et al., in press). AutoTutor is moderately successful in tracking these emotions on the basis of the dialogue history, facial expressions of the learner, posture of the learner, and speech intonation. AutoTutor detects the correct emotion approximately half of the time over and above a base rate of 0, approximately the same likelihood that human experts would classify these emotions.
How might an emotion-sensitive AutoTutor help learning? As discussed earlier, AutoTutor should respond very differently when the learner is in these different emotional states. AutoTutor needs to deliver razzle-dazzle when the student is bored, hints when the student is frustrated, encouragement when experiencing delight, typical responses when the student is engaged (a state of flow), and so on. Confusion is an interesting affective state because that is when thinking is most pronounced in the learner. AutoTutor should not cut off such productive thinking, but how much confusion should occur before AutoTutor needs to step in with some productive direction? These are all questions for future research. No one has investigated the relationship between learning and an emotion-sensitive tutor.

**CLOSING COMMENTS**

Animated pedagogical agents have the potential to have a revolutionary impact on education. Researchers can systematically manipulate and constrain how these agents produce messages, give feedback, and respond to learners in a tailored intelligent fashion. This is nearly impossible to achieve when researchers attempt to train human instructors how to systematically deliver the curriculum, messages, strategies, and style of a classroom or tutorial interaction. Humans fall prey to regressing to natural habits of teaching and conversing whereas computer agents can mind the theoretical pedagogy. We are at a special point in history when these computer agents go a long distance in simulating human speech, content, and emotions. It would be wise to capitalize on this technology rather than to have the knee-jerk reaction that computers are not human. Designers of these learning environments are indeed very human.

**AUTHOR NOTES**

The research on AutoTutor was supported by the National Science Foundation (SBR 9720314, REC 0106965, REC 0126265, ITR 0325428, REESE 0633918), the Institute of Education Sciences (R305H050169), and the DoD 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, DoD, or ONR. The Tutoring Research Group (TRG) is an interdisciplinary research team comprised of researchers from psychology, computer science, physics, and education (visit http://www.autotutor.org). Requests for reprints should be sent to Art Graesser, Department of Psychology, 202 Psychology Building, University of Memphis, Memphis, TN 38152-3230, a-graesser@memphis.edu.
REFERENCES


