Two Heads May Be Better Than One: Learning From Computer Agents in Conversational Trialogues

ARTHUR C. GRAESSER

University of Memphis

CAROL M. FORSYTH

Educational Testing Service

BLAIR A. LEHMAN

Educational Testing Service

Background: *Pedagogical agents are computerized talking heads or embodied animated avatars that help students learn by performing actions and holding conversations with the students in natural language. Dialogues occur between a tutor agent and the student in the case of AutoTutor and other intelligent tutoring systems with natural-language conversation. The agents are adaptive to the students' actions, verbal contributions, and, in some systems, their emotions (such as boredom, confusion, and frustration).*

Focus of Study: *This paper explores several designs of trialogues (two agents interacting with a human student) that have been productively implemented for particular students, subject matters, and depths of learning. The two agents take on different roles, but often serve as peers and tutors. There are different trialogue designs that address different pedagogical goals for different classes of students. For example, students can (a) observe vicariously two agents interacting, (b) converse with a tutor agent while a peer agent periodically chimes in, or (c) teach a peer agent while a tutor rescues a problematic interaction. In addition, agents can argue with each other over issues and ask what the human student thinks about the argument.*

Research Design: *Trialogues have been developed for systematic experimental investigations in several studies that measure student impressions, learning gains from pretest to post-test on objective tests, and both cognitive and affective states during learning. The studies compare conditions with different pedagogical principles underlying the trialogues in order to assess the impact of these principles on student impressions, learning, emotions, and other psychological measures. Discourse analyses are performed on the language and actions in the log files in order to assess their impacts on psychological measures.*

Teachers College Record Volume 119, 030302, March 2017, 20 pages
Copyright © by Teachers College, Columbia University
0161-4681

Recommendations: *Tests of these agent-based systems have shown improvements in learning gains and systematic influences on student emotions. In the future, researchers need to conduct more research to empirically evaluate the psychological impact of different trialogue designs on psychological measures. These trialogue designs range from scripted interactions between agents being observed by the student, to the student helping a fellow peer agent, to the student resolving an argument between two agents. The central question is whether the learning experiences and outcomes show improvement over typical human-computer dialogues (i.e., one human and one tutor agent) and conventional pedagogical interventions.*

Pedagogical agents are computerized talking heads or embodied animated avatars that generate speech, actions, facial expressions, and gestures. The agents relevant to this article are adaptive to the actions, language, and sometimes the emotions of the student, as opposed to providing rigid, choreographed displays of spoken language and action. Adaptive pedagogical agents have been developed to serve as substitutes for humans who range in expertise from peers to subject-matter experts with pedagogical strategies. Agents can guide the student on what to do next, deliver didactic instruction, hold collaborative conversations, and model ideal behavior, strategies, reflections, and social interactions.

Pedagogical agents have become increasingly popular in contemporary adaptive learning environments. Examples of adaptive pedagogical agents that have successfully improved student learning are AutoTutor (Graesser et al., 2004, 2012; Nye, Graesser, & Hu, 2014), DeepTutor (Rus, D'Mello, Graesser, & Hu, 2013), GuruTutor (Olney et al., 2012), Betty's Brain (Biswas, Jeong, Kinnebrew, Sulcer, & Roscoe, 2010), Tactical Language and Culture System (Johnson & Valente, 2008), Coach Mike (Lane, Noren, Auerbach, Birth, & Swartout, 2011), iDRIVE (Gholson et al., 2009), iSTART (Jackson & McNamara, 2013; McNamara, O'Reilly, Best, & Ozuru, 2006), Crystal Island (Rowe, Shores, Mott, & Lester, 2010), Operation ARA (Halpern et al., 2012; Millis et al., 2011), and My Science Tutor (Ward et al., 2013). These systems have covered topics in STEM (physics, biology, computer literacy), reading comprehension, scientific reasoning, and other domains and skills. At this point in the science, researchers are investigating the conditions in which particular designs of pedagogical agents significantly improve learning and/or motivation for particular categories of students on particular subject matters and skills.

The most common design of agent interaction consists of a dialogue, in which the human student interacts with only one agent. The agent can be either a peer (approximately the same level of proficiency as the human), a student agent with lower proficiency (so that the learner can teach the agent), or an expert tutor agent. AutoTutor is a pedagogical agent that simulates the dialogue moves of human tutors in addition to ideal pedagogical strategies (Graesser et al., 2004, 2012; Graesser, Jeon, & Dufty, 2008; Nye et al., 2014). Simulating a tutor is a sensible first design for agents because human tutoring is known to be a very effective method for improving student learning and motivation. Meta-analyses that compare tutoring to classroom teaching and other suitable comparison conditions report effect sizes between $\sigma = 0.20$ and $\sigma = 1.00$ (Cohen, Kulik & Kulik, 1982; Graesser, D'Mello, & Cade, 2011; VanLehn, 2011).

Researchers have investigated why human tutoring is so effective in helping students learn (Cade, Copeland, Person, & D'Mello, 2008; Chi, Siler, Yamouchi, Jeong, & Hausmann, 2001; Graesser, D'Mello, & Person, 2009; Graesser, Person, & Magliano, 1995). Tutor effectiveness does not simply consist of lecturing the student, but rather arises from the tutor's attempts to get the student to construct answers and generate solutions to problems. Tutor effectiveness also cannot be attributed to (a) a fine-grained diagnosis of what the student knows, (b) high shared knowledge between the tutor and student, or (c) consistent accurate feedback to the student. Why so? Because human tutors have limited abilities to diagnose the students' knowledge, and their shared knowledge is minimal. Instead, most human tutors follow a systematic conversational structure that is called *expectation and misconception-tailored (EMT) dialogue* (Graesser et al., 2008, 2012). That is, human tutors anticipate particular correct answers (called expectations) and particular misconceptions when they ask the students challenging questions or problems and trace their reasoning. As a particular student articulates answers over multiple conversational turns, the student's contributions are compared with the expectations and misconceptions, and the tutor thereby forms an approximate model of what the student knows. The tutor gives feedback to the student that depends on how well the contributions match the expectations or misconceptions. The tutors produce *dialogue moves* to encourage the students to generate content and eventually cover the expectations. Misconceptions are corrected by the tutor when expressed by the students.

Listed below are tutor dialogue moves that frequently occur in the EMT dialogues in AutoTutor and in most human tutoring sessions.

- **Main question or problem:** This is a challenging question or problem that the tutor is trying to help the student answer. It may take 5 to over 100 conversational turns to answer collaboratively.
- **Short feedback:** The feedback to a student's contribution is either *positive* ("yes," "correct," head nod), *negative* ("no," "almost," head shake, long pause, frown), or *neutral* ("uh-huh," "OK").

- **Pumps:** The tutor gives nondirective pumps ("Anything else?" "Tell me more") to get the student to do the talking or to take some action.
- **Hints:** The tutor gives hints to get the students to do the talking or take action, but directs the students along some conceptual path. The hints vary from generic statements or questions ("What about X?" "Why?") to speech acts that nudge the student toward a particular answer. Hints promote active student learning within the boundaries of relevant material.
- **Prompts:** The tutor asks a very leading question in order to get the student to articulate a particular word or phrase. Sometimes students say very little, so these prompts are needed to get the student to say something specific.
- **Prompt completions:** The tutor expresses the correct completion of a prompt.
- **Assertions:** The tutor expresses a fact or state of affairs.
- **Summaries:** The tutor gives a recap of the answer to the main question or solution to the problem.
- **Mini-lectures:** The tutor expresses didactic content on a particular topic.
- **Corrections:** The tutor corrects an error or misconception of the student.
- **Answers:** The tutor answers a question asked by the student.
- **Off-topic comment:** The tutor expresses information unrelated or tangentially related to the subject matter.

These dialogue moves differ on the extent to which the student versus the tutor supplies the expectation content. For example the tutor supplies progressively more of the expectation information as one goes from pump to hint to prompt to assertion to summaries. According to the principle of active student learning, there should be a greater onus on the student, rather than the tutor, in supplying the expectation information. Indeed, tutorial dialogues with more knowledgeable students have a higher proportion of tutor pumps and hints, requiring greater student input, rather than prompts and assertions that provide more information from the tutor (Jackson & Graesser, 2006).

The EMT dialogue of AutoTutor helps students learn challenging material. AutoTutor shows learning gains of approximately 0.80 σ(standard deviation units) compared with reading a textbook for an equivalent amount of time (Graesser et al., 2012; Nye et al., 2014). This assessment of learning is based on over 20 experiments in the areas of computer literacy (Graesser et al., 2004), Newtonian physics (VanLehn et al., 2007), and scientific reasoning (Kopp, Britt, Millis, & Graesser, 2012). Approximately a dozen measures of learning have been collected in these assessments, including multiple-choice questions, essay quality when students attempt to answer challenging questions, a cloze task that has students fill in missing words of texts that articulate explanatory reasoning on the subject matter, and performance on problems that require problem-solving. AutoTutor is most impressive for increasing learning gains on measures of deeper rather than shallow learning. Shallow knowledge and learning include mastery and memory of simple facts, rules, and procedures. Deeper knowledge and learning require generating inferences, integrating information, scrutinizing the validity of claims, reasoning, and problem-solving. It is more difficult for students to achieve deeper learning, so they turn to tutors for help.

PEDAGOGICAL AGENTS IN TRIALOGUES

Our contention is that adding a second agent to form a trialogue will have pedagogical benefits in the design of tutoring systems with pedagogical agents (Graesser, Li, & Forsyth, 2014). Multiple agents have been incorporated in many learning environments with agents, such as Betty's Brain (Biswas et al., 2010), Tactical Language and Culture System (Johnson & Valente, 2008), iDRIVE (Gholson et al., 2009), iSTART (Jackson & McNamara, 2013; McNamara et al., 2006), and Operation ARA (Forsyth et al., 2012; Halpern et al., 2012; Millis et al., 2011).

We have recently explored several designs of trialogues in order to acquire a systematic understanding of how trialogues can be productively implemented for particular students, subject matters, and depths of learning. Listed below are some trialogue designs that researchers have investigated recently. We start with *vicarious learning* designs, where learners learn by observing others, in this case pedagogical agents:

a. **Vicarious learning with human observer:** Two agents interact and model ideal behavior, answers to questions, or reasoning. The two agents can be peers, tutors/experts, or a mixture.

b. **Vicarious learning with limited human participation:** The same type of interaction as Design 1, except that the agents occasionally turn to the human and ask a prompt question, with a yes/no or single-word answer. The prompt questions foster human engagement and assessment of human comprehension.

c. **Tutor agent interacting with human and student agent:** There is a tutorial dialogue with the human, but the student agent periodically contributes and receives feedback. Negative short feedback can be given to the student agent on bad answers (the agent takes the heat), whereas similar answers by the human student receive neutral feedback.

d. **Expert agent staging a competition between the human and a peer agent:** There is a competitive game (with points scored) between the human and peer agent, with the competition guided by the expert agent.

e. **Human teaches/helps a student agent with facilitation from the tutor agent:** This is a "teachable agent" design (learning by teaching), with the tutor agent rescuing problematic interactions. More advanced students might particularly benefit from this design.

f. **Human interacts with two peer agents that vary in proficiency:** The peer agents can vary in knowledge and skills. In assessment contexts, the computer can track whether the human corrects an incorrect contribution by a peer, correctly answers a peer's question, and takes initiative in guiding the exchange.

g. **Human interacts with two agents expressing contradictions, arguments, or different views:** The discrepancies between agents stimulate cognitive disagreement, confusion, and potentially deeper learning.

The vicarious learning designs (1 and 2) are appropriate for students with limited knowledge, skills, and actions, as well as training for shallow learning rather than deep learning, as discussed earlier. Agent designs 5 and 7 are appropriate for more capable students and deeper learning. The agent designs are also differentially suited to the motivation and emotions of the learner. For example, design 4 is motivating for students by virtue of the game competition, design 3 minimizes negative feedback to the student, and design 7 elicits confusion (a major predictor of deep learning: see D'Mello, Pekrun, Lehman, & Graesser, 2014; Lehman et al., 2013). Finally, the trialogue designs allow assessment of the student's proficiency to varying degrees. Design 1 provides no on-line assessment, whereas assessment in design 2 is minimal, and that in 3–7 is substantial. Performance is assessed by comparing student actions and verbal input with the expectations and misconceptions.

Now that we have outlined some trialogue designs, we briefly describe three research projects that implement them. We start with Operation ARIES/ARA, a serious game that attempts to train students in critical thinking about scientific methodology (Forsyth et al., 2012; Halpern et al., 2012; Millis et al., 2011). We then turn to some experiments on scientific reasoning that address the affective state of the student to facilitate deeper learning. Specifically, we attempt to induce cognitive disequilibrium and confusion by trialogue design pattern 7. The central question in this research is how the disequilibrium and confusion play a role in deeper learning. We end with a description of using agent trialogues to help struggling adult readers learn how to read better in our Center for the Study of Adult Literacy. The first two examples have proved successful in helping students learn, whereas the third awaits future empirical testing.

OPERATION ARIES! AND ARA ON SCIENTIFIC REASONING

The trialogues in this section are based on an instructional game called Operation ARIES! (Millis et al., 2011), which had a version commercialized on an experimental basis by Pearson Education as Operation ARA (Halpern et al., 2012). ARIES is an acronym for Acquiring Research Investigative and Evaluative Skills, whereas ARA is an acronym for Acquiring Research Acumen. It takes approximately 20 hours to complete ARIES and 10 hours to complete ARA.

The game teaches high school and college students how to critically evaluate aspects of scientific investigations, such as the need for control groups, random assignment, and operational definitions and the difference between correlation and causation. In the game phases under consideration in this article, the two agents and human student collaboratively answer difficult questions based on a text they have read or critique case studies of scientific research with respect to scientific methodology. Three types of trialogues are implemented in ARIES and launched under specific conditions. They are (a) vicarious learning with human participation (trialogue design 2), (b) tutorial learning (designs 3 or 4), and (c) learning through teaching (design 5). The type of trialogue that occurs for a particular core concept is based on the level of knowledge exhibited by the player earlier in the game. Low knowledge launches vicarious learning trialogues, intermediate knowledge launches tutorial learning, and high knowledge launches learning through teaching.

The conversation below illustrates many of the dialogue moves of Expectation and Misconception Tailored (EMT) tutoring, but in this case there is a trialogue in Operation ARA rather than a dialogue as in AutoTutor. The conversation includes examples of many of the tutor's moves in Operation ARA: main question, short feedback (negative, neutral, and positive), correction, pump, prompt, correct prompt completion, hint, and summary. Tom is the name of the human student, Dr. Quinn is the teacher agent, and Glass is a peer agent. The information in italics and brackets annotates some of the dialogue move categories. Sometimes Glass gives incorrect answers, whereas Dr. Quinn has the ground truth. The trialogue encourages Tom to fill in information, answer questions, and resolve disagreements. Therefore, the human is an active participant (trialogue design 2) rather than merely a passive vicarious observer (design 1).

Dr. Quinn: Let's talk about this some more.

Dr. Quinn: In a true experiment, why do you think random assignment is important when placing participants in an experimental or control group? *[Main question]*

Tom: it makes sure everyone has the same chance of being in the experiment

Glass: I think so too, Tom.

Glass: Random assignment is important because it makes it so that everyone in the same population has an equal chance of being in the study.

Dr. Quinn: That is incorrect, Glass. *[Negative feedback]*

Dr. Quinn: You are thinking about random selection. This is about random assignment. *[Correction]* Try saying more about this topic. *[Pump]*

Glass: Okay, so I know that random assignment ensures that the control and experimental groups are similar, but I'm not sure why. I think this has to do with the fact that the goal of the study is to show that an independent variable has an effect on a dependent variable. How is this related? *[Peer hint]* Tom, can you help me understand?

Tom: It makes it so that the variability is spread out

Dr. Quinn: OK. *[Neutral feedback]*

Dr. Quinn: With random assignment, subjects have an equal chance of being in any group, and this makes it so that all of the groups should show a fairly similar distribution of differences in subject what? *[Prompt]*

Tom: characteristics

Dr. Quinn: Great! *[Positive feedback]*

Dr. Quinn: A similar distribution of differences in subject characteristics. *[Correct prompt completion]*

> **Dr. Quinn**: Think about this.
>
> **Dr. Quinn:** In a true experiment, we know that random assignment helps evenly distribute participant characteristics across the groups. How does this help us understand the effects of the manipulation on the outcome of a study? *[Hint]*
>
> **Tom:** we know the effect is statistically significant
>
> **Glass:** No, I don't think that is correct. [*Negative feedback*]
>
> **Glass:** I think that random assignment is important because it shows us that the independent variable had the effect on the dependent variable rather than the characteristics of the participants in the study.
>
> **Dr. Quinn**: In summary, random assignment across experimental and control groups . . . . *[Summary]*

The trialogues in ARIES and ARA have been shown to help students learn (Forsyth et al., 2012; Halpern et al., 2012; Millis et al., 2011), but we are still exploring what characteristics of the conversational interaction account for the learning. We do know that AutoTutor dialogues and trialogues are substantially better than reading texts on the same content for an equivalent amount of time (Graesser et al., 2014; Kopp et al., 2012; VanLehn et al., 2007). However, we need to better understand the underlying mechanisms and the discourse components of ARA that account for the learning gains.

As in AutoTutor, ARIES/ARA has an EMT dialogue mechanism in which the students' verbal contributions are compared with expectations and misconceptions for an answer. The goal of the trialogue is to help the player articulate a specific expectation (sentence) in the exchange, such as, "A scientific hypothesis must have a prediction that can be tested." If the student answers it correctly, then the trialogue is finished, and the next expectation is considered or the next question is asked. The student receives a high performance score of 100% if the student immediately articulates the expectation. If the student's answer is incomplete, then the tutor agent gives a hint, such as, "What about testing a hypothesis?" A correct answer to the hint yields a score of 67%. If the student answer is still incomplete, then the tutor agent gives a leading prompt question, such as, "What is tested when there is a scientific hypothesis?" with the hope that the student fills in the word "prediction" and thereby gets partial credit of 33%. If the student is still incorrect, then the peer agent can interject and give the correct answer, after which the tutor agent gives positive feedback to the peer agent (whereas the human learner receives 0% credit).

Instead of giving negative feedback to the incorrect human student, the tutor gives positive feedback to the correct peer agent. This promotes politeness and avoids face-threatening negative feedback to the human. Thus, in addition to optimizing motivation there is a metric of scaling the student's performance during the course of conversational scaffolding.

The trialogues can be designed with additional discourse approaches (i.e., tutor discourse moves) to press the envelope of learning, social interaction, and rigorous assessment. A few examples illustrate such attempts. After the trialogue covers the expectation through multiple turns, an agent can *request a summary* from the student, e.g., "Could you summarize what a hypothesis is?" From the standpoint of active learning, it is better for the student to provide a summary, rather than an agent. An agent can *request a verification* from the student to verify whether a statement is true or false, e.g., "Do hypotheses require a prediction?" At various points during the trialogue, an agent can *barge in* and interrupt the thread of exchange between the student and another agent with a question or other speech act. In the case of *student agent echoing*, the agent expresses something similar to the student and receives accurate feedback from the tutor agent. Another approach to handling uncertainty in what the student is saying is for one of the agents to *request clarification*, e.g., "I don't understand," "Could you rephrase that?" "Could you be more precise?" Indeed, agents are allowed to express that they don't understand, just as people do. There can be a scheme of *praising the human and blaming the peer student agent*. In this fashion, the peer agent gets the brunt of negative attributions, whereas the human student is provided with positive feedback. Yet another approach is *eliciting cognitive disequilibrium*. Cognitive disequilibrium is planted by manipulating whether or not the tutor agent and the student agent contradict each other during the trialogue or express claims that are incorrect (D'Mello et al., 2014; Lehman et al., 2013). This approach is addressed in the next section.

CONTRADICTIONS, CONFUSION, AND DEEP LEARNING

Humans experience a variety of emotions and affective states during learning, particularly when the material is difficult. This has motivated us to conduct studies that track the emotions of students during interactions with AutoTutor (D'Mello & Graesser, 2012; Graesser & D'Mello, 2012) and other intelligent tutoring systems (R. S. Baker, D'Mello, Rodrigo, & Graesser, 2010; Forsyth et al., 2013). The common emotions that students experience during a learning session of 1–2 hours are boredom, engagement/flow, frustration, confusion, delight, and surprise, in a wide range of learning environments. Of these emotions, the emotion that best predicts learning at deeper levels is confusion, a cognitive-affective state associated with thought and deliberation (Craig, Graesser, Sullins, & Gholson, 2004; D'Mello & Graesser, 2012; D'Mello et al., 2014; Graesser & D'Mello, 2012). We have investigated a cognitive disequilibrium framework that integrates a number of psychological processes: confusion (and other learning-centered emotions), question-asking (inquiry), deliberative thought, and deeper learning. Cognitive disequilibrium is a state that occurs when people face obstacles to goals, interruptions, contradictions, incongruities, anomalies, impasses, uncertainty, and salient contrasts (D'Mello & Graesser, 2012; Festinger, 1957; Mandler, 1999; Piaget, 1952). Initially the person experiences various emotions when beset with cognitive disequilibrium, most notably confusion and surprise (D'Mello & Graesser, 2012; D'Mello et al., 2014; Graesser & D'Mello, 2012; Lehman, D'Mello, & Graesser, 2012; Lehman et al., 2013). This disequilibrium elicits question-asking and other forms of inquiry (Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005; Graesser & McMahen, 1993; Otero & Graesser, 2001), such as eye movements, physical exploration of the environment, and social interaction. The person engages in problem-solving, reasoning, and other thoughtful cognitive activities in an attempt to resolve the impasse and restore cognitive equilibrium. One consequence is deeper learning.

Trialogues can be designed to manipulate cognitive disequilibrium (see trialogue design 7). This is accomplished by having the tutor agent and the student agent contradict each other or disagree during the trialogue. The ARIES/ARA case studies on potentially flawed research served as the materials in a program of research to investigate cognitive disequilibrium, confusion, and deep learning with trialogues (D'Mello et al., 2014; Lehman et al., 2012, 2013). More specifically, the tutor agent and student agent engaged in a short exchange about (a) whether there was a flaw in a study and (b) the nature of the flaw if there *was* a flaw. There were four variations in how these trialogues could occur. In the *True-True* control condition, the tutor agent expressed a correct assertion, and the student agent agreed with the tutor. In the *True-False* condition, the tutor expressed a correct assertion, but the student agent disagreed by expressing an incorrect assertion. The *False-True* condition was the flip side, with the tutor expressing a false assertion, whereas the *False-False* condition had both the tutor and the student agent agreeing on incorrect information.

The central question was whether the contradictions would plant confusion and subsequent reasoning at deeper levels, which in turn would improve learning. We assumed the following approximate sequence of events:

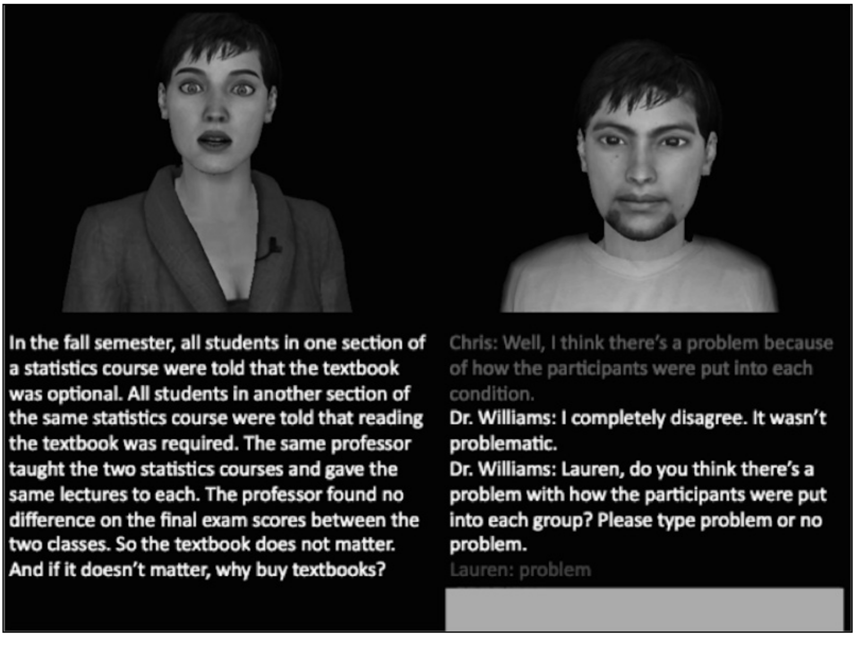
contradiction → cognitive disequilibrium → confusion →

inquiry & deep reasoning → deep learning

We were uncertain about the timing of the middle three processes, but there was some theoretical foundation for this ordering. Additional research is clearly needed to pinpoint the order of processes, the time duration of the processes, and their causal status.

In order to convey the manipulation of cognitive disequilibrium more concretely, Figure 1 presents a screen shot of the tutor and student agents, a case study, and the beginning of a trialogue that launches the disagreement. The tutor agent (Dr. Williams) is on the left, the student agent (Chris) is on the right, and the human student is Lauren. The case study (on the bottom left) describes an experiment that assesses the impact of a textbook on learning in a statistics course. There is an issue of the adequacy of the experimental design with respect to participant assignment to groups. The disagreement between Chris and Dr. Williams is on the bottom right, and Lauren is asked for her opinion about whether there is a problem in how participants were assigned to groups.

Figure 1. Screenshot of two agents disagreeing about an experiment



In one measure of confusion, the agents asked the student questions after several particular points of agent contradiction in the conversation (i.e., agent trialogue design 2). For example, the agents turned to the human and asked, “Do you think there's a problem with how the participants were put into each group?” The tutor agent (Dr. Williams) states that she believes that the assignment was not problematic, whereas the peer student agent (Chris) disagrees and believes that the assignment was flawed. The human student (Lauren) is asked to decide which agent is correct. So the human is put in cognitive disequilibrium en route to a decision. The quality of the responses to these forced-choice questions is presumably a signal of confusion. A confused student would be expected to respond incorrectly or in a manner inconsistent with previous responses (e.g., oscillating between positions posed by the agents). The experience of confusion would then potentially stimulate thinking, reasoning, and learning.

The data indeed confirmed that the contradictions had an impact on students' answers to these forced-choice questions immediately following a contradiction. Response correctness showed the following order for the four trialogue conditions: True-True > True-False > False-True > False-False. These findings indicated that learners typically agreed with the agents when the agents agreed (True-True, False-False), but were often confused when there was a contradiction between the two agents (True-False, False-True). It was also found that when the agents disagreed, learners shifted their responses between agreeing with the tutor agent versus the student agent more frequently compared to the conditions in which the agents agreed. This confusion was predicted to elicit deeper reasoning and problem solving, which should be helpful to learning. Interestingly, there was also some evidence that disequilibrium and/or confusion caused more learning at deeper levels of mastery, as reflected on a transfer test of scientific reasoning. Specifically, experimental conditions with agent contradictions often produced higher performance on assessments (multiple-choice questions, flaw identification tasks) that tapped deep levels of comprehension compared with performance in the True-True condition. But this deep learning occurred only if students were confused by the contradictions during training. These results are consistent with the hypothesis that there may be a causal relationship between cognitive disequilibrium and deep learning, with confusion playing a moderating role on the effect of the contradictions on learning.

It is illuminating that the False-False condition did not engender much uncertainty and confusion. The students pretty much accepted what the tutor and student agents expressed when they agreed, even if the claims were false. Thus, the perceived power dynamics placed the judgment of the two agents above the human's judgment (in those potential instances when the human would have had a different judgment). The results could have turned out very differently. Specifically, the claims of the two agents could have clashed with the reasoning and belief system of the human. Interestingly, this alternative possibility was found to occur with high prior knowledge college students who participated in this study, but did not occur overall. This result is compatible with models that predict that it takes a large amount of knowledge about a subject matter before students can detect what they don't know (knowledge gaps) (Miyake & Norman, 1979), false information (Rapp, 2008), and contradictory information (L. Baker, 1985). Students need to take a strategic, skeptical, critical stance if they are not fortified with sufficient subject-matter knowledge. Thus, there is a need for the two agents to directly contradict each other in a conversation before the students experience an appreciable amount of uncertainty and confusion.

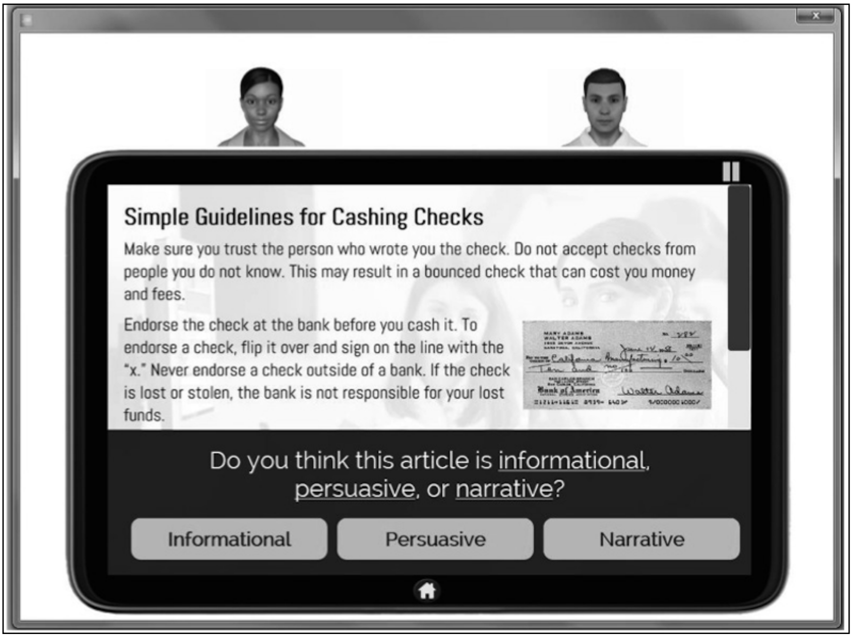
We also suspect that the contradictory statements need to be contiguous in time for the contradiction to be detected. The contradiction is likely to be missed if one agent makes a claim, and then another agent makes a contradictory claim 10 minutes later. This is compatible with research in text comprehension that has shown that the contradictory claims must be co-present in working memory before they get noticed (L. Baker, 1985), unless there is a high amount of world knowledge. It is also compatible with the observation that it is difficult for students to integrate information from multiple texts and spot contradictions (Rouet, 2006; Wiley et al., 2009) unless there is a high amount of world knowledge. The agent trialogues can bring these contradictions under focus and thereby finesse the cognitive disequilibrium and confusion that lead to deeper learning.

TRIALOGUES FOR IMPROVING ADULT LITERACY

Agent trialogues are currently being developed in computer interventions for the Center for the Study of Adult Literacy (CSAL, http://csal.gsu.edu/content/homepage). The goal is to help adults with low literacy acquire strategies for comprehending text at multiple levels of language and discourse. There are several comprehension strategies covered in the intervention, including predicting features of text genre, acquiring vocabulary from context, clarifying the explicit text through questioning, elaborating the text through inferences, and identifying text structures. Agents with speech are particularly appropriate for this computer intervention because these adults have limited reading comprehension abilities.

Figure 2 shows a snapshot of the interface for trialogues in the CSAL project, with a teacher agent (Cristina, top left), a student agent (Jordan, top right), and the human who interacts with the agents. The text describes some guidelines for cashing checks, a very practical text for the adult reader. The task is for the adult to make a judgment whether the text genre is informational, persuasive, or narrative. The learner is expected to click on one of the three options at the bottom under the printed question (which is also spoken by the teacher agent). There is considerable multimedia in the CSAL trialogues to keep the adult reader's attention. Most of the student input consists of clicking on response options and elements in the text because the adult's writing is even more limited than his or her reading. However, in later lessons the learner is expected to type in words, phrases, and eventually 1–2 sentences during a conversational turn.

Figure 2. Screenshot of conversational agents in a trialogue to help adult learners read



Most of the trialogue designs are implemented in the CSAL intervention. The agents guide the human through the experience, model good practice and interactions between agents, and give feedback and explanations for correct or incorrect answers. In some lessons there is competition between the human and student agent, and scores are kept in the game competition. When considering all of the lessons in CSAL, the agents and humans can serve many functions: teacher, helper, collaborator, ideal model, judge, and critic. Assessments of performance can be extracted from the trialogue interactions, based on the accuracy of the adults' decisions, clicks on options, and verbal contributions.

FINAL COMMENTS

This article has illustrated some designs of trialogues that can facilitate learning, manipulate emotions, and provide a foundation for assessment. Trialogues provide some advantages over dialogues. For example, it would not be possible to model social interactions, stage competitions, and manipulate cognitive disequilibrium without two agents. As expressed in the title of this article, two heads may sometimes be better than one.

At this point there still needs to be more systematic research on the conditions in which particular trialogue designs are effective in facilitating learning. For example, there are few studies that support the contention that vicarious learning (trialogue designs 1 and 2) is best for low-ability students, tutoring (trialogue design 3) is best for intermediate-ability students, and learning by teaching (trialogue design 5) is best for high-ability students. Aside from the gradient of learner abilities, there is the gradient of their emotional and motivational states. Additional investigations are needed to determine when competition (trialogue design 4) increases or decreases engagement, and when cognitive disequilibrium (trialogue design 7) induces frustration and disengagement rather than confusion and deep learning.

Researchers are just beginning to explore trialogues for rigorous assessments of learning and other psychological constructs. Educational Testing Service is incorporating trialogues in conversation-based assessments for many proficiencies, including virtual-world environments in assessments of English Language Learning (ELL) and science (Zapata-Rivera, Jackson, & Katz, 2015). In the area of ELL there can be assessments of reading, writing, listening, and speaking in one integrated environment. Trialogues may potentially be used to disentangle student aptitudes in multiple domains, such as assessing ELL and mathematics at the same time. This venture may aid researchers to discover whether a potential deficit in ELL has implications for low scores in mathematics when quantitative skills may not actually be the issue. This is quite a different world than the decades of assessment in the past.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (SBR 9720314, REC 0106965, REC 0126265, ITR 0325428, REESE 0633918, ALT-0834847, DRK-12-0918409, 1108845), the Institute of Education Sciences (R305H050169, R305B070349, R305A080589, R305A080594, R305A090528, R305A100875, R305C120001), the Army Research Laboratory (W911INF-12-2-0030), and the Office of Naval Research (N00014-12-C-0643, N00014-16-C-3027). 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.

REFERENCES

Baker, L. (1985). Differences in standards used by college students to evaluate their comprehension of expository prose. *Reading Research Quarterly, 20,* 298–313.

Baker, R. S., D'Mello, S. K., Rodrigo, M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68,* 223–241.

Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology-Enhanced Learning, 5,* 123–152.

Cade, W., Copeland, J. Person, N., & D'Mello, S. K. (2008). Dialogue modes in expert tutoring. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems* (pp. 470–479). Berlin, Germany: Springer-Verlag.

Chi, M. T. H., Siler, S., Yamauchi, T., Jeong, H., & Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science, 25,* 471–534.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19,* 237–248.

Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning. *Journal of Educational Media, 29,* 241–250.

D'Mello, S. K., & Graesser, A. C. (2012). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems, 2,* 1–39.

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. C. (2014). Confusion can be beneficial for learning. *Learning and Instruction, 29,* 153–170.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford University Press.

Forsyth, C. M., Graesser, A. C. Pavlik, P., Cai, Z., Butler, H., Halpern, D. F., & Millis, K. (2013). Operation ARIES! methods, mystery and mixed models: Discourse features predict affect in a serious game. *Journal of Educational Data Mining, 5,* 147–189.

Forsyth, C. M., Pavlik, P., Graesser, A. C., Cai, Z., Germany, M, Millis, K., . . . Dolan, R. (2012). Learning gains for core concepts in a serious game on scientific reasoning. In K. Yacef, O. Zaïane, H. Hershkovitz, M. Yudelson, & J. Stamper (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 172–175). Chania, Greece: International Educational Data Mining.

Gholson, B., Witherspoon, A., Morgan, B., Brittingham, J. K., Coles, R., Graesser, A. C., . . . Craig, S. D. (2009). Exploring the deep-level reasoning questions effect during vicarious learning among eighth to eleventh graders in the domains of computer literacy and Newtonian physics. *Instructional Science, 37,* 487–493.

Graesser, A. C., & D'Mello, S. (2012). Emotions during the learning of difficult material. In. B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 57, pp. 183–225). San Diego, CA: Academic Press.

Graesser, A. C., D'Mello, S. K., & Cade, W. (2011). Instruction based on tutoring. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 408–426). New York, NY: Routledge Press.

Graesser, A. C., D'Mello, S. K., Hu. X., Cai, Z., Olney, A., & Morgan, B. (2012). AutoTutor. In P. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation, and resolution* (pp. 169–187). Hershey, PA: IGI Global.

Graesser, A. C., D'Mello, S., & Person, N. K. (2009). Metaknowledge in tutoring. In D. J. Hacker, J. Donlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 361–382). New York, NY: Taylor & Francis.

Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes, 45*, 298–322.

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science, 23,* 374–380.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers, 36*, 180–193.

Graesser, A. C., Lu, S., Olde, B. A., Cooper-Pye, E., & Whitten, S. (2005). Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down. *Memory and Cognition, 33*, 1235–1247.

Graesser, A. C., & McMahen, C. L. (1993). Anomalous information triggers questions when adults solve problems and comprehend stories. *Journal of Educational Psychology, 85*, 136–151.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*, 495–522.

Halpern, D. F., Millis, K., Graesser, A. C., Butler, H., Forsyth, C., & Cai, Z. (2012). Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. *Thinking Skills and Creativity, 7,* 93–100.

Jackson, G. T., & Graesser, A. C. (2006). Applications of human tutorial dialog in AutoTutor: An intelligent tutoring system. *Revista Signos, 39*, 31–48.

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology, 105*, 1036–1049.

Johnson, L. W., & Valente, A. (2008). Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. In M. Goker & K. Haigh (Eds.), *Proceedings of the Twentieth Conference on Innovative Applications of Artificial Intelligence* (pp. 1632–1639). Menlo Park, CA: AAAI Press.

Kopp, K., Britt, A., Millis, K., & Graesser, A. C. (2012). Improving the efficiency of dialogue in tutoring. *Learning and Instruction, 22*, 320–330.

Lane, H. C., Noren, D., Auerbach, D., Birch, M., & Swartout, W. (2011). Intelligent tutoring goes to the museum in the big city: A pedagogical agent for informal science education. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *International Journal of Artificial Intelligence in Education* (pp. 155–162). Heidelberg, Germany: Springer.

Lehman, B., D'Mello, S. K., & Graesser, A. C. (2012). Confusion and complex learning during interactions with computer learning environments. *Internet and Higher Education, 15*(3), 184–194.

Lehman, B., D'Mello, S. K., Strain, A., Mills, C., Gross, M., Dobbins, A., . . . Graesser, A. C. (2013). Inducing and tracking confusion with contradictions during complex learning. *International Journal of Artificial Intelligence in Education, 22*, 85–105.

Mandler, G. (1999). Emotion. In B. M. Bly & D. E. Rumelhart (Eds.), *Cognitive science* (Handbook of perception and cognition, 2nd ed., pp. 367–384). San Diego, CA: Academic Press.

McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research, 34,* 147–171.

Millis, K, Forsyth, C., Butler, H., Wallace, P., Graesser, A. C., & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, & J. Lakhmi (Eds.*), Serious games and edutainment applications* (pp.169–196). London, UK: Springer-Verlag.

Miyake, N., & Norman, D. A. (1979). To ask a question one must know enough to know what is not known. *Journal of Verbal Learning and Verbal Behavior, 18,* 357–364.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education, 24,* 427–469.

Olney, A., D'Mello, S. K., Person, N., Cade, W., Hays, P., Williams, C., . . . Graesser, A. C. (2012). Guru: A computer tutor that models expert human tutors. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of Intelligent Tutoring Systems (ITS) 2012* (pp. 256–261). Berlin, Germany: Springer.

Otero, J., & Graesser, A. C. (2001). PREG: Elements of a model of question asking. *Cognition & Instruction*, *19*, 143–175.

Piaget, J. (1952). *The origins of intelligence.* New York: International University Press.

Rapp, D. N. (2008). How do readers handle incorrect information during reading? *Memory & Cognition, 36*, 688–701.

Rouet, J.-F. (2006). *The skills of document use.* Mahwah, NJ: Erlbaum.

Rowe, J., Shores, L. R., Mott, B., & Lester, J. (2010). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, *21*(1–2), 166–177.

Rus, V., D'Mello, S., Hu, X., & Graesser, A. C. (2013). Recent advances in intelligent systems with conversational dialogue. *AI Magazine, 34,* 42–54.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. *Educational Psychologist*, *46,* 197–221.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*, 3–62.

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., & Weston, T. (2013). My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology, 105*, 1115–1125.

Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. *American Educational Research Journal, 46,* 1060–1106.

Zapata-Rivera, D., Jackson, T., & Katz, I. R. (2015). Authoring conversation-based assessment scenarios. In. R. Sottilare, A. C. Graesser, X. Hu, & K. Brawner (Eds.), *Design recommendations for intelligent tutoring systems: Authoring tools* (Vol. 3, pp. 191–200). Orlando, FL: Army Research Laboratory.

ARTHUR C. GRAESSER is Distinguished University Professor of Interdisciplinary Research in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis and is an Honorary Research Fellow in the Oxford University Center for Educational Assessment at the University of Oxford. His primary research interests are in cognitive science, discourse processing, computational linguistics, and the learning sciences. He has developed automated tutoring systems with conversational agents (such as AutoTutor and Operation ARA) and automated text analysis systems (Coh-Metrix, QUAID). Recent publications include "Deeper Learning with Advances in Discourse Science and Technology," in *Policy Insights from Behavioral and Brain Sciences*, and "Intelligent Tutoring Systems, Serious Games, and the Generalized Intelligent Framework for Tutoring (GIFT)," in *Using Games and Simulation for Teaching and Assessment.*

CAROL M. FORSYTH is now Associate Research Scientist in the Cognitive, Accessibility, & Technology Sciences Center at Educational Testing Service. She received her Ph.D. in Cognitive Psychology and Cognitive Science Certificate from the University of Memphis in 2014. Her research interests include intelligent tutoring systems, epistemic games, and discourse processes during natural language conversations for tutoring and assessment. Recent publications include "Operation ARIES! Methods, Mystery and Mixed Models: Discourse Features Predict Affect in a Serious Game," in *Journal of Educational Data Mining*, and "Discourse Comprehension," in the *Oxford Handbook of Cognitive Psychology.*

BLAIR A. LEHMAN is now an Associate Research Scientist in the Cognitive, Accessibility, & Technology Sciences Center at Educational Testing Service. She received her Ph.D. in Cognitive Psychology and Cognitive Science Certificate from the University of Memphis in 2014. Her research interests include human and computer tutoring, emotions during learning, and natural language conversation for tutoring and assessment. Recent publications include "Inducing and Tracking Confusion with Contradictions During Complex Learning," in *International Journal of Artificial Intelligence in Education*, and "Confusion Can Be Beneficial for Learning," in *Learning & Instruction.*