International Journal of Artificial Intelligence in Education 22 (2013) 85–105
DOI 10.3233/JAI-130025
IOS Press

Inducing and Tracking Confusion with Contradictions during Complex Learning

Blair Lehman, *University of Memphis, Memphis, TN, USA*
Sidney D'Mello, *University of Notre Dame, Notre Dame, USA, sdmello@nd.edu*
Amber Strain, *University of Memphis, Memphis, TN, USA*
Caitlin Mills, *University of Notre Dame, Notre Dame, USA*
Melissa Gross, *University of Memphis, Memphis, TN, USA*
Allyson Dobbins, *University of Memphis, Memphis, TN, USA*
Patricia Wallace, *Northern Illinois University, DeKalb, IL, USA*
Keith Millis, *Northern Illinois University, DeKalb, IL, USA*
Art Graesser, *University of Memphis, Memphis, TN, USA*

Abstract. Cognitive disequilibrium and its affiliated affective state of confusion have been found to positively correlate with learning, presumably due to the effortful cognitive activities that accompany their experience. Although confusion naturally occurs in several learning contexts, we hypothesize that it can be induced and scaffolded to increase learning opportunities. We addressed the possibility of confusion induction in a study where learners engaged in trialogues on research methods concepts with animated tutor and student agents. Confusion was induced by staging disagreements and contradictions between the animated agents, and then inviting the (human) learners to provide their opinions. Self-reports of confusion indicated that the contradictions were successful at inducing confusion in the minds of the learners. A second, more objective, method of tracking learners' confusion consisted of analyzing learners' performance on forced-choice questions that were embedded after contradictions. This measure was also found to be revealing of learners' underlying confusion. The contradictions alone did not result in enhanced learning gains. However, when confusion had been successfully induced, learners who were presented with contradictions did show improved learning compared to a no-contradiction control. Theoretical and applied implications along with possible future directions are discussed.

Keywords. Confusion, cognitive disequilibrium, contradictions, affect, tutoring, intelligent tutoring systems, learning

INTRODUCTION

Research over the last decade has made major advances in understanding the connection between affect and learning. One of the major findings from this line of research is that cognition and affect are not distinct or opposing constructs. When learners struggle they can experience a host of negative affective states, such as confusion, frustration, and even anger in more extreme cases. However, when learners conquer a challenge or achieve a major goal they may experience happiness or joy. How learners affectively respond to challenges and achievements can influence learners' motivation and engagement. A better understanding of both the cognitive and affective experiences of learners is needed to fully understand the learning process and to maximize learning gains.

A range of affective states can be experienced during complex learning. In-depth analyses of relatively short learning sessions (30 minutes to 1.5 hours) have revealed a distinct set of *learning-centered affective states* (Arroyo, et al., 2009; Burleson & Picard, 2007; Chaffar, Derbali, & Frasson, 2009; Conati & Maclaren, 2009; D'Mello, Craig, Fike, & Graesser, 2009; Forbes-Riley & Litman, 2011; Lehman, Matthews, D'Mello, & Person, 2008; Robison, McQuiggan, & Lester, 2009; Rodrigo & Baker, 2011b). The learning-centered affective states include anxiety, boredom, confusion, curiosity, delight, engagement/flow, frustration, happiness, and surprise (Calvo & D'Mello, 2011; Rodrigo & Baker, 2011a). Learning-centered affective states have also been found to be differentially related to learning outcomes. Experiences of confusion and flow were positively correlated with learning outcomes, whereas boredom and frustration were negatively correlated with learning during interactions with an intelligent tutoring system (ITS) called AutoTutor (Craig, Graesser, Sullins, & Gholson, 2004; Graesser, Chipman, King, McDaniel, & D'Mello, 2007).

However, there are still open questions about the connection between affect and learning despite the advances that have been made in recent research. There is the question of how learning environments can incorporate affect into instructional strategies to maximize learning. One approach is to respond when affective states naturally arise during learning. For example, a learning environment can ask learners to self-explain when they are confused or provide hints and other scaffolds when learners falter (D'Mello et al., 2010). A second approach is to induce affective states that facilitate learning. We have adopted this latter strategy in the present paper. Our focus is on confusion, a state that many intuitively believe is negative and harmful to learning, but in fact has been found to be positively correlated with learning under certain circumstances, as will be elaborated below (Barth & Funke, 2010; Craig, et al., 2004; Graesser, et al., 2007; D'Mello & Graesser, 2011). We investigated a learning environment that experimentally induces confusion as a method to increase learning opportunities and positively impact learning outcomes.

Theory and previous research

Confusion is considered to be an epistemic or knowledge affective state (Pekrun & Stephens, 2012; Silvia, 2010) that occurs when learners reach an impasse, are confronted with a contradiction, anomaly, or system breakdown, and are uncertain about what to do next (Brown & VanLehn, 1980; Caroll & Kay, 1988; VanLehn et al., 2003). Learners are placed in a state of cognitive disequilibrium and experience confusion when these discrepant events occur (Bjork & Linn, 2006; Festinger, 1957; Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005; Piaget, 1952). Confusion and cognitive disequilibrium can trigger cognitive activities such as reflection and problem solving that can be beneficial for learning. In fact, increased experiences of confusion during learning have been linked to deeper comprehension (Craig et al., 2004; D'Mello & Graesser, 2011; Graesser et al., 2007; Lehman, D'Mello, & Graesser, 2012). For example, in an analysis of over 100 hours of human-human tutorial dialogues, VanLehn et al. (2003) reported that comprehension of physics concepts was rare when learners did not reach an impasse, irrespective of the quality of explanations provided by the tutor. However, it is unlikely that the mere occurrence of confusion leads to increased learning. It is presumably not confusion itself but rather the effects of confusion resolution processes that lead to increased learning (D'Mello & Graesser, 2012; VanLehn et al., 2003).

Although confusion can provide opportunities for learning, it is not the case that all experiences of confusion are beneficial for learning. Learners can choose to not engage in confusion resolution and would then be unlikely to reach a deeper level of understanding. In fact, in Chinn and Brewer's (1993) classification scheme for how people respond to contradictory information and anomalous data, four of the seven response patterns involve generally dismissing the presence of new, problematic information. Even when learners attend to contradictions, there are still cases in which experiences of confusion can be detrimental to learning. Hopeless, or persistent, confusion occurs when learners are unable to resolve experiences of confusion (D'Mello & Graesser, 2012). This should be contrasted with productive confusion, which can

immediately or eventually be resolved. An event that is too challenging for the learner could create hopeless confusion, but with the proper scaffolds and support the learner can experience productive confusion for the same event.

The question arises, how can learning environments take advantage of the benefits of confusion, while also avoiding any potential pitfalls, to maximize learning? We propose that the benefits of confusion can only be leveraged in a learning environment if three fundamental conditions are met: (1) the learning environment has events that *induce* confusion, (2) the learning environment can detect and *track* the associated confusion, and (3) the learning environment *regulates* (scaffolds) the confusion in a way that maximizes learning.

Past research would indicate that many events during complex learning naturally create confusion (Bhatt et al., 2004; Craig et al., 2004; D'Mello, Craig, Sullins, & Graesser, 2006; Graesser et al., 2007; Lehman et al., 2008; Litman & Forbes-Riley, 2004; Pon-Barry, Schultz, Bratt, Clark, & Peters, 2006; Porayska-Pomsta, Mavrikis, & Pain, 2008; Robison et al., 2009). For example, an analysis of tutorial dialogue during interactions with AutoTutor revealed that confusion occurred when learners attempted to answer challenging questions, when the tutor provided scaffolding with less direct questions (e.g., hints, pumps), and when the tutor provided negative feedback (D'Mello et al., 2006; D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008). Thus, one approach to scaffold confusion is to take advantage of these naturally occurring experiences of confusion during learning.

There are currently ITSs that adaptively respond to learner emotions when they naturally occur during learning (Arroyo, et al., 2009; Burleson & Picard, 2007; Chaffar et al., 2009; Conati & Maclaren, 2009; D'Mello et al., 2009; Forbes-Riley & Litman, 2011; Robison et al., 2009). Forbes-Riley and Litman (2011) have created a version of ITSPOKE, a spoken dialogue ITS, that adaptively responds to learner uncertainty, a state that is similar to experiences of confusion. In a Wizard of Oz experimental design, Forbes-Riley and Litman tested two methods of adaptively responding to learner uncertainty. There were four identified impasse severities based on answer correctness and uncertainty: correct + certainty (no impasse), correct + uncertainty, incorrect + uncertainty, and incorrect + certainty (most severe impasse). In one method a human selected the response that ITSPOKE deployed and responded to all impasses in the same manner (simple), whereas in the other method a human differentially selected responses to impasses based on severity (complex). The simple method in which all impasses were treated the same was the most effective for learning when compared to a version of ITSPOKE that did not adapt to learner uncertainty; however, learners preferred the more complex method.

Another approach is to proactively induce the state of confusion in learners, rather than wait for confusion to naturally occur and then respond. In two experiments, D'Mello and Graesser (in review) adopted this approach by inducing confusion through the presentation of device breakdowns. In these experiments, participants were first presented with an illustrated text of a device (e.g., a cylinder lock) and given 1.5 to 2 minutes to study how the device worked. Participants were then presented with the same illustrated text and an additional prompt. This additional prompt described some type of breakdown that had occurred with the device (experimental condition). The cylinder lock, for example, had the following breakdown: "A person puts the key into the lock and turns the lock but the bolt doesn't move." Participants were then asked to try to determine why the device was not functioning. The control condition involved either re-reading the original illustrated text (Experiment 1) or focusing on a key component of the device while re-reading (Experiment 2). Breakdown scenarios were found to induce greater levels of confusion compared to the control condition. In addition, participants who resolved or partially-resolved their confusion performed better on a device comprehension test than those who did not resolve their confusion.

In the present study we conducted a preliminary investigation of another confusion induction method within a dialogue-based learning environment. We attempted to induce confusion through the presentation of contradictory information. Contradictory information, if attended to, can trigger confusion and cognitive disequilibrium in the mind of learners (Carroll & Kay, 1988; Chan, Burtis, & Bereiter, 1997; Chinn &

Brewer, 1993; VanLehn et al., 2003). In the current learning environment, the human learner engaged in a trialogue (three-party conversation) with two animated pedagogical agents. One agent served the role of tutor and the other served as a peer student agent. The two pedagogical agents served as the medium through which confusion was induced over the course of learning research methods concepts, such as designing and evaluating research studies (Halpern, 2003; Roth et al., 2006).

Two research questions guided the present study. First, can confusion be experimentally induced through the presentation of contradictory information? In other words, will confusion be induced if one agent presents accurate information and the other agent presents inaccurate information? Second, what are the indicators that can be used to detect and track confusion in a learning environment? A third issue relevant to the present study is how learning is impacted by confusion induction. Although experiences of confusion can trigger cognitive activities that are beneficial for learning, we do not expect any impressive learning gains in the present study. This is because this initial investigation did not provide any intervention or scaffolding to help learners manage and resolve their confusion.

METHOD

Participants

Participants were 32 undergraduate students from a mid-south university in the US who participated for course credit. Data from one participant was discarded due to experimenter error. There were 21 females and 11 males in the sample. Participants' age ranged from 18 to 43 years old ($M = 21.7$, $SD = 6.10$). Sixty-seven percent of participants were African-American, 30% were Caucasian, and 3% were Hispanic. Prior coursework in research methods was not required for participation. Ninety percent of participants had not taken a research methods or a statistics course prior to participation.

Learning activity

The central learning activity consisted of critiquing research case studies to determine whether they exhibit sound scientific methodology or have particular methodological flaws. Participants engaged in a trialogue (three-party conversation) with two animated pedagogical agents (tutor and student) to evaluate the scientific merits of the case studies. Note that student agent refers to an animated agent; the human learner is referred to as participant or learner for the remainder of the paper. Critical evaluation of case studies involves scientific reasoning skills such as stating hypotheses, identifying dependent and independent variables, isolating potential confounds in designs, and determining if data support predictions (Halpern, 2003; Roth et al., 2006). During the evaluation of a case study, each agent presented its opinion on the scientific merits of the case study and invited the participant to intervene. For example, in one trialogue the tutor agent asserted that the study was flawed whereas the student agent disagreed and asserted that the study was flawless. After both agents presented their respective opinions, the tutor agent then asked the participant whether he or she believed that the study contained a flaw. The trialogue continued in this manner, with the discussion becoming increasingly more specific about the nature of the flaw in the case study (discussed further below). Altogether, participants completed eight trialogues with the two agents.

Manipulation

We experimentally induced confusion with a contradictory information manipulation. Contradictions were introduced during trialogues that identified flaws in case studies. This manipulation was achieved by having

the tutor and student agents stage a disagreement on a concept and eventually invite the participant to intervene. The contradiction was expected to trigger conflict and force the participant to reflect, deliberate, and decide which opinion had more scientific merit. When participants were invited to intervene, they had to decide if they agreed with the tutor agent, the student agent, both agents, or neither of the agents.

There were four contradictory information conditions. In the *true-true* condition, the tutor agent presented a correct opinion and the student agent agreed with the tutor; this is the no-contradiction control. In the *true-false* condition, the tutor presented a correct opinion and the student agent disagreed by presenting an incorrect opinion. In contrast, it was the student agent who provided the correct opinion and the tutor agent who disagreed with an incorrect opinion in the *false-true* condition. Finally in the *false-false* condition, the tutor agent provided an incorrect opinion and the student agent agreed. It should be noted that all misleading information was corrected at the end of each of the trialogues and participants were fully debriefed at the end of the experiment.

Design

The experiment had a within-subjects design with four conditions: *true-true, true-false, false-true,* and *false-false*. Participants completed two trialogues in each of the four conditions with a different research methods concept discussed in each session (8 in all). The eight research methods concepts were construct validity, control groups, correlational studies, experimenter bias, generalizability, measure quality, random assignment, and replication. Each concept had an associated research case study that might or might not have been flawed in one significant aspect. Half the studies for a given participant contained flaws and the other half were flawless. Order of conditions and concepts and assignment of concepts to conditions was counterbalanced across participants with a Graeco-Latin Square. The presence of flaws was also counterbalanced across concepts and conditions.

Knowledge tests

Research methods knowledge on the eight concepts covered in the trialogues was tested before and after trialogues with two knowledge tests (pretest and posttest, respectively). Each test had 24 multiple-choice questions with three questions per concept. There were three types of test items: definition, function, and example. The definition questions were relatively shallow, the function questions targeted the utility or function of each concept, and the example questions involved applications of the concepts. Random assignment, for example, was assessed with the following questions: "Random assignment refers to __" (definition), "Random assignment is important because __" (function), and "Which study most likely did not use random assignment?" (example) (see Appendix A for additional examples). There were two alternate test versions and assignment of test version was counterbalanced across participants for pretest and posttest.

Procedure

Participants were individually tested over a three-hour session. The experiment occurred over two phases: (1) knowledge assessments and trialogues and (2) retrospective affect judgment protocol.

Knowledge assessments and trialogues

First, participants signed an informed consent and were seated in front of a computer console with a widescreen (21.5") monitor with 1920 × 1080 resolution and an integrated webcam. Next, participants completed the pretest and read a short introductory text on research methods. The introductory text provided

participants with a broad overview of the research methods terminology that was discussed during the trialogues.

Participants were instructed to put on a set of headphones after reading the introductory text. The tutor and student agents introduced themselves, discussed their roles, discussed the importance of developing research methods knowledge and skills, and described the learning activity. Participants were not informed that the agents may disagree or provide inaccurate opinions prior to the learning sessions. Participants then began the first of eight trialogues. The structure of each trialogue is described in detail below. Each trialogue lasted for 6.50 minutes ($SD = 1.20$), and the eight trialogues took an average total of 53.5 minutes ($SD = 6.78$) to complete. After completing all eight trialogues, participants completed the multiple-choice posttest. Participants were fully debriefed at the end of the experiment.

In each trialogue, participants discussed the case study for one of the research methods concepts with the tutor and student agents. The trialogue interface is shown in Fig. 1 and consisted of (A) the tutor agent, (B) the student agent, (C) a description of the case study, (D) a text transcript of the dialogue history, and (E) a text box for participants to enter and submit their responses. The tutor and student agents delivered the content of their utterances via synthesized speech, while the participant typed his or her responses.

The trialogue for each case study involved four multi-turn trials. For example, in Table 1, turns four through eight represent one multi-turn trial. The following activities occurred in each trial: (1) one agent provided an opinion on an aspect of the study, (2) the second agent either concurred with that opinion or disagreed by providing an alternate opinion, (3) the tutor agent asked the participant for his or her opinion via a forced-choice question, (4) the participant provided his or her opinion, and (5) the participant was asked to explain his or her opinion (third and fourth trials only). For each forced-choice question the participant was required to give a valid response. The trialogue would only proceed if the participant had responded with one of the two response options presented by the tutor agent.

This cycle was repeated in each trial, with each trial becoming increasingly more specific about the scientific merits of the study. As an example, consider the trialogue in Table 1 (*true-false* condition) in which the tutor agent (Dr. W), student agent (Chris), and participant (Bob) discussed a study that was improperly replicated. Trial 1 generally considered if participants would change their behavior based on the results of the study (turns 1–3), while Trial 2 addressed whether or not the methodology of the study was problematic (turns 4–8). Trial 3 was specifically concerned with whether the replication conducted was good or bad (turns 9–13). Finally, Trial 4 directly addressed *why* the replication of the study was or was not flawed (turns 14–17). Participants were also required to provide an explanation about their response to the forced-choice questions in Trials 3 and 4 because the discussion was more specific for these trials. For example, after the participant responded "different" in Trial 4, the tutor agent would say: "Bob, explain to me why you think that" (not shown in Table 1).

The trialogue structure was identical across all four conditions with a small exception. The only differences occurred when the agents stated their opinions about the scientific merit of the case study based on the contradictory information condition. This difference manifested as the agent either stating that the concept was or was not flawed in the study being discussed. In the trialogue in Table 1, for example, Chris (student agent) stated "I don't think there's anything wrong with how they did this study" (turn 5). The example in Table 1 is from the *True-False* condition in which Chris stated an incorrect opinion. In the *False-True* condition, on the other hand, Chris stated a correct opinion. Thus, in the *False-True* condition Chris stated "I think there's something wrong with how they did this study" (turn 5). The two versions of this opinion only differed on the agent's opinion, but did not differ on the content presented.

All contradictory and false information was corrected after Trial 4. This consisted of the tutor agent asserting whether the study contained a flaw and then providing an accurate evaluation of the study (turn 18). Finally, the tutor agent provided a brief explanation of how the study could have been improved (turn 20).

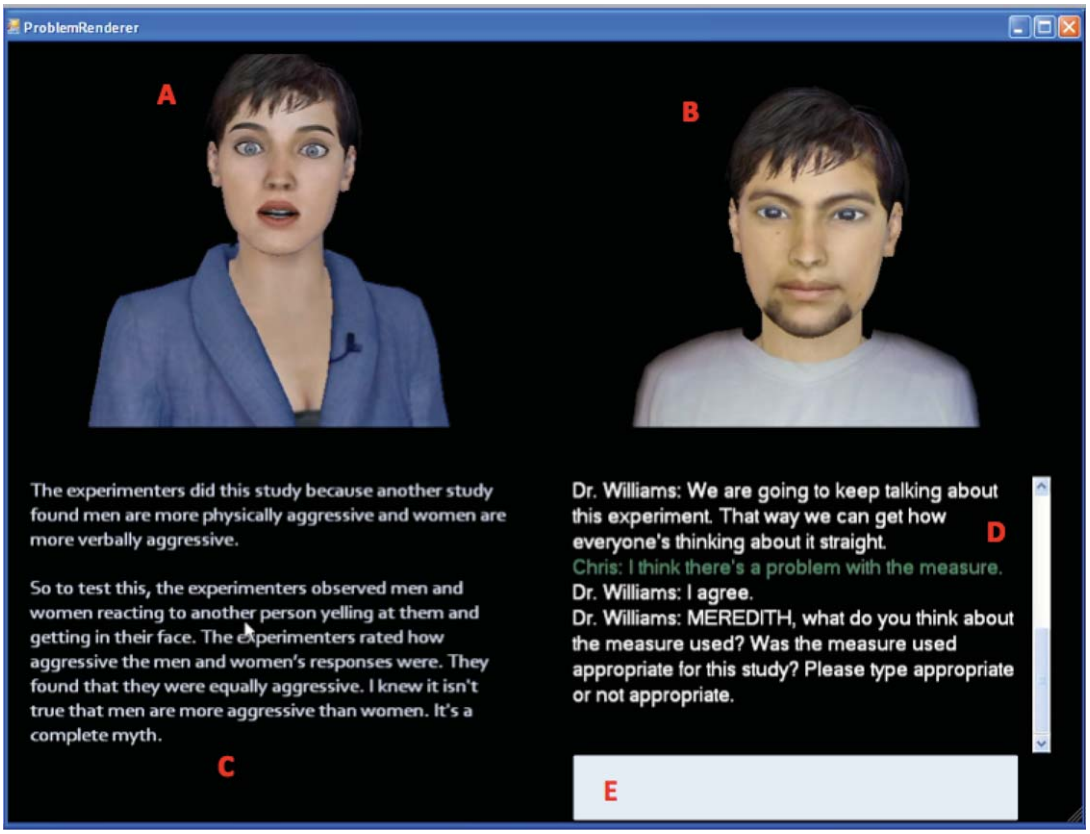

Fig. 1. Screen shot of learning interface.

Three streams of information were recorded as participants completed the trialogues. First, a video of the participant's face was captured using a webcam that was integrated into the computer monitor. The webcam also recorded all audio generated during the interaction. Second, a video of the participant's screen was recorded using a commercially available screen capture program called Camtasia Studio™. Third, a variety of interaction parameters were automatically recorded in log files. These parameters included the participant's responses (typed responses and response times) and the current state of the interaction (e.g., pretest vs. trialogue).

Retrospective affect judgment protocol

Participants completed a retrospective affect judgment protocol (Graesser et al., 2006) immediately after finishing the posttest. Videos of participants' faces and screens were synchronized and participants made affect ratings over the course of viewing these two videos. Figure 2 shows the interface used for the retrospective affect judgment protocol. Participants were provided with a list of affective states (anxiety, boredom, confusion, curiosity, delight, engagement/flow, frustration, surprise, and neutral) with definitions. The list of affective states was motivated by previous research on affective states during learning with technology (D'Mello & Graesser, in press; Rodrigo & Baker, 2011a).

The affective states were explicitly defined before participants made their judgments and participants had access to the definitions while making their judgments. Anxiety was defined as being nervous, uneasy, apprehensive, or worried. Boredom was defined as being weary or restless through lack of interest.

Table 1
Excerpt of trialogue from *true-false* condition

Turn	Speaker	Dialogue
A study done with university undergraduates showed that people who drank 6 cups of non-alcoholic beer self-reported levels of intoxication close to that of people who drank 6 cups of actual beer. Here's how the study was done. A professor ran this study with his upper level psych class. There were 100 students in the class over 21, so 50 drank the real beer and 50 drank the non-alcoholic beer. The professor wanted to make sure that his results were real so he decided to replicate the study two weeks later. He did the same exact study with the same class, making sure to randomly assign people to conditions. He found the same results again. So he concluded that people could feel just as drunk from drinking non-alcoholic beer as drinking alcoholic beer. <*Case study*>		
1	Dr. W	So we've been talking and Chris thinks that there are no problems with this study, but I think there's a problem. <*Tutor & student assertion*>
2	Dr. W	Bob, do you think people need alcohol to feel drunk? Type alcohol or no alcohol. <*Forced-choice question*>
3	Bob	no alcohol < *Response*>
4	Dr. W	Since Chris and I disagree, let's talk some more about this experiment. <*Advance dialogue*>
5	Chris	I don't think there's anything wrong with how they did this study. <*Assertion*>
6	Dr. W	I don't think so, I disagree. <*Assertion*>
7	Dr. W	Bob, do you think this was a good methodology? Type good or faulty. <*Forced-choice question*>
8	Bob	faulty <*Response*>
9	Dr. W	Ok. Chris, was there anything particularly problematic with either study? <*Advance dialogue*>
10	Chris	Well let me think about this. I think the replication was done right. <*Assertion*>
11	Dr. W	I disagree. <*Assertion*>
12	Dr. W	Bob, was this a good or bad replication? Type good or bad. <*Forced-choice question*>
13	Bob	bad <*Response*>
14	Chris	You know, I think using the same people for both studies makes it a good replication. <*Assertion*>
15	Dr. W	No, using the same population doesn't strengthen his findings. <*Assertion*>
16	Dr. W	Bob, do you think that he should've used the same population for his replication study? Type same or different. <*Forced-choice question*>
17	Bob	different <*Response*>
18	Dr. W	Well now that we've discussed this study I'm even more certain that I was right. Using the exact same population doesn't make a good replication. <*Correct assertion*>
19	Chris	What would've been a better replication study? <*Advance dialogue*>
20	Dr. W	Well, for example, he could've run the same experiment in another one of his classes or with professors. If he had found the same results in one or all of those studies then he could be really confident about his findings. <*Summary*>

Confusion was defined as a noticeable lack of understanding and being unsure about how to proceed. Curiosity was defined as a desire to acquire more knowledge or learn the material more deeply. Delight was defined as a high degree of satisfaction. Engagement/flow was defined as a state of interest that results from involvement in an activity. Frustration was defined as dissatisfaction or annoyance from being stuck. Surprise was defined as a state of wonder or amazement, especially from the unexpected. Finally, neutral

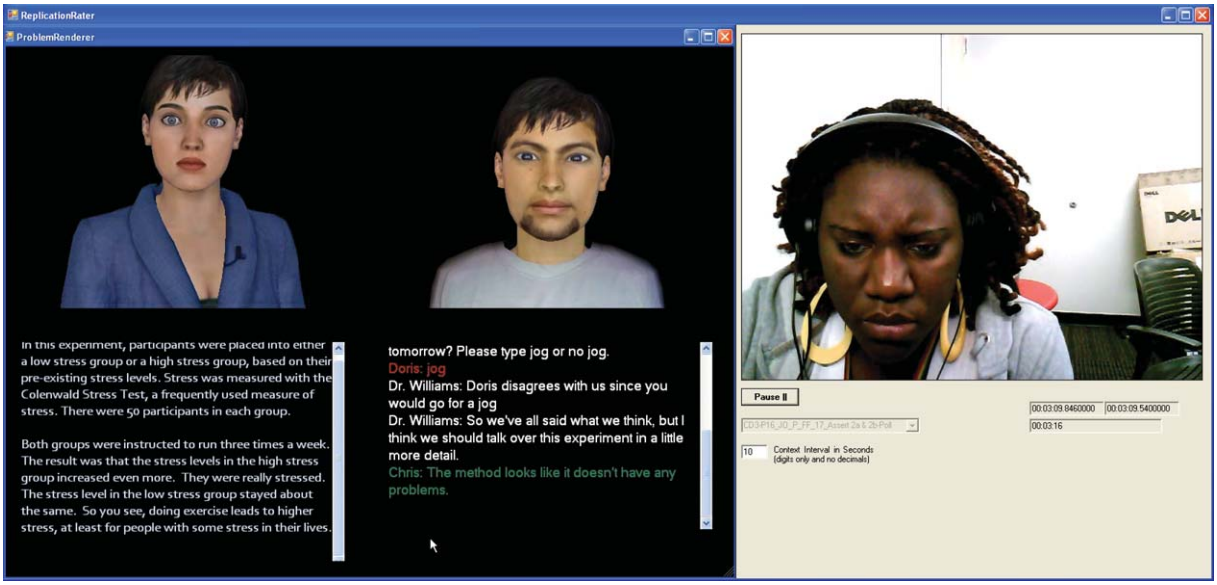

Fig. 2. Screenshot of retrospective affect judgment protocol interface.

was defined as having no apparent emotion or feeling. It should be noted that the affect judgments were not based on these definitions alone, but on the combination of videos of participants' faces, contextual cues via the screen capture, the definitions of the affective states, and participants' recent memories of the interaction.

Affect judgments occurred at 13 pre-specified points in each learning session (104 in all). The majority of the pre-specified points focused on the contradictory information events in the trialogues. Participants were required to report their affective state (i.e., presence of one of the eight affective states or neutral) after both agents provided their opinions (turns 1, 6, 11, and 15 in Table 1), after the forced-choice questions were posed (turns 2, 7, 12, and 16), and after participants were asked to explain their responses in Trials 3 and 4 (not shown in Table 1). Participants also reported their affective state after reading the case study, at the end of the learning session when the tutor agent stated whether the study contained a flaw, and after the tutor agent explained the scientific merits of the study (turns 18 and 20 in Table 1). In addition to these pre-specified points, participants were able to manually pause the videos and provide affect judgments at any time (voluntary judgments).

RESULTS AND DISCUSSION

The analyses were guided by three research objectives: (a) evaluate the presentation of contradictory information as a method of confusion induction, (b) investigate methods of tracking confusion during learning, and (c) explore the impact of confusion induction on learning outcomes. For each analysis, the experimental conditions (*true-false*, *false-true*, and *false-false*) were compared to the no-contradiction control condition (*true-true*). There were four primary dependent measures in the present analyses: (1) self-reported affect obtained via the retrospective affect judgment protocol, (2) accuracy of responses to forced-choice questions following contradictions, (3) shifts in response accuracy across trials (e.g., Trial 1 to 2), and (4) performance on the multiple-choice posttest.

A mixed-effects modeling approach was used to conduct a majority of the analyses in the present study. Mixed-effects modeling is the recommended approach for this type of data due to the repeated measurements and nested structure of the data (trials nested within case studies, case studies aligned with conditions) (Pinheiro & Bates, 2000). Mixed-effects models include a combination of fixed and random effects and can be used to assess the influence of the fixed effects on dependent variables after accounting for any extraneous random effects. The *lme4* package in R (Bates & Maechler, 2010) was used to perform the requisite computations.

Linear or logistic models were constructed on the basis of whether the dependent variable was continuous or binary, respectively. The random effects were *participant* (31 levels), *case study* (8 levels), and *order* (order of presentation of case study). *Condition* was a four-level (*true-true*, *true-false*, *false-true*, and *false-false*) categorical fixed effect.

The comparisons reported in this paper focus on the a priori comparison of each experimental condition to the no-contradiction control, so the *true-true* condition was set as the reference group in all of the models. One-tailed tests were used for significance testing when the hypothesis specified the direction of the effect. However, two-tailed tests were used when no a priori predictions were made. Specific hypotheses for each analysis conducted are discussed in their respective sub-sections of the results section. All significance testing was conducted with an alpha level of 0.05.

Self-reported affect

The retrospective affect judgment procedure yielded 3224 judgments at the pre-specified points (fixed judgments) and 237 voluntary judgments provided by the participants. Due to the small number of voluntary judgments, they were combined with the fixed judgments, thereby yielding a total of 3461 affect judgments. Nine mixed-effects logistic regressions that detected the presence (coded as 1) or absence (coded as 0) of each affective state were constructed. The unit of analysis was an individual affect judgment, so there were 3461 cases in the data set.

For self-reported affect, we predicted that more confusion would be reported in the experimental conditions compared to the control condition. However, we did not have specific predictions about the other affective states being self-reported. Comparisons for these other self-reported affective states were tested using two-tailed tests. Significant[1] models were discovered for all affective states, with the exception of anxiety and frustration: boredom ($\chi^2(3) = 14.9$, $p = .002$), confusion ($\chi^2(3) = 5.866$, $p = .055$, curiosity ($\chi^2(3) = 9.16$, $p = .027$), delight ($\chi^2(3) = 10.6$, $p = .014$), engagement/flow ($\chi^2(3) = 9.83$, $p = .020$), neutral ($\chi^2(3) = 8.95$, $p = .030$), and surprise ($\chi^2(3) = 8.15$, $p = .043$).

The coefficients for the models along with the mean proportional occurrence of each affective state are presented in Table 2. An analysis of the model coefficients indicated that participants self-reported significantly more confusion when in the *true-false* condition than in the *true-true* condition. The difference between the estimates (i.e., B values) for this comparison was 0.558, so participants were $e^{0.558}$ or 1.75 times more likely to report confusion when in the *true-false* condition compared to the *true-true* condition. Participants were also significantly less likely to report boredom and delight ($p = .103$) when in the *true-false* condition compared to the *true-true* condition.

There were no significant differences in participants' self-reported confusion when in the *false-true* condition compared to the *true-true* condition. However, participants were significantly less likely to report experiences of curiosity and neutral when in the *false-true* condition compared to the *true-true* condition. When participants were in the *false-false* condition, they were more likely to report higher levels of confusion

[1] Significance of a mixed-effects logistic model is evaluated by comparing the mixed-model (fixed + random effects) to a random model (random effects only) with a likelihood ratio test.

Table 2
Proportional occurrence of affective states

	Proportional occurrence				Coefficient B and odds ratio (exp(B))					
Affect	*Tr-Tr*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*	*Tr-Fl*		*Fl-Tr*		*Fl-Fl*	
Anxiety	.004	.005	.004	.002	.252	(1.29)	−.023	(.977)	−.441	(.643)
Boredom	.256	.232	.290	.306	**−.256**	(.774)	.197	(1.22)	**.323**	(1.38)
Confusion	.036	.062	.043	.049	**.558**	(1.74)	.221	(1.25)	.354	(1.42)
Curiosity	.057	.052	.035	.032	−.092	(.912)	**−.589**	(.555)	**−.708**	(.493)
Delight	.043	.030	.051	.043	−.586	(.557)	.562	(1.75)	.204	(1.23)
Engaged	.289	.312	.280	.261	.143	(1.15)	−.141	(.868)	**−.300**	(.741)
Frustration	.021	.030	.038	.039	.411	(1.51)	.590	(1.80)	.639	(1.89)
Neutral	.280	.255	.241	.234	−.182	(.834)	**−.302**	(.739)	**−.378**	(.685)
Surprise	.014	.022	.018	.034	.379	(1.46)	.147	(1.16)	**.890**	(2.44)

*Note*s. Tr: True; Fl: False. Tr-Tr was the reference group for each model, hence, coefficients for this condition are not shown in the table. Bolded cells refer to significant effects at $p < .05$.

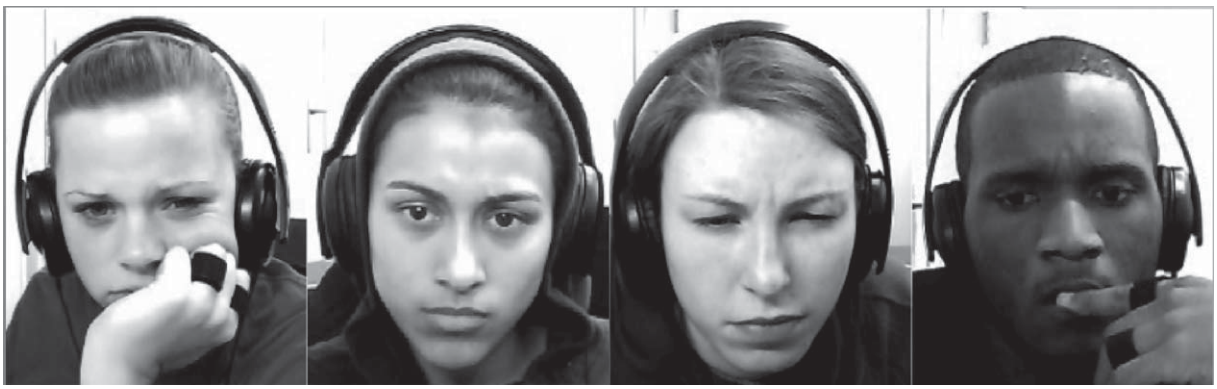

Fig. 3. Examples of confused faces from participants.

($p = .083$) and surprise and lower levels of curiosity, engagement/flow, and neutral compared to the *true-true* condition.

These findings suggest that contradictions between agents can induce some confusion in participants. Figure 3 shows four examples of participants in the experimental conditions with confused expressions who also reported experiencing confusion. The success of confusion induction, however, does appear to be tempered by who (tutor vs. student) takes the correct vs. incorrect position.

Responses to forced-choice questions

Self-reports are one viable method to track confusion. However, this measure is limited by the participant's sensitivity and willingness to report their confusion levels. A more subtle and objective measure of confusion is to assess participant responses to forced-choice questions following contradictions. We hypothesized that participants' overall performance on these forced-choice questions would be reduced due to uncertainty about how to proceed with evaluating the case study. Thus, participant confusion could be inferred from the

Table 3
Proportion of forced-choice questions correctly answered

	Proportion correct				Coefficient B and odds ratio (exp(B))					
Trial	*Tr-Tr*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*	*Tr-Fl*		*Fl-Tr*		*Fl-Fl*	
Trial 1	.581	.532	.500	.403	−.206	(.814)	−.341	(.711)	**−.752**	(.471)
Trial 2	.871	.565	.484	.274	**−1.73**	(.177)	**−1.80**	(.165)	**−2.78**	(.062)
Trial 3	.855	.677	.339	.290	**−1.10**	(.333)	**−2.66**	(.067)	**−2.91**	(.054)
Trial 4	.726	.613	.468	.419	−.516	(.597)	**−1.07**	(.343)	**−1.27**	(.281)

Notes. Tr: True; Fl: False; Tr-Tr was the reference group for each model, hence, coefficients for this condition are not shown in the table. Bolded cells refer to significant effects at $p < .05$.

accuracy of their responses to these forced-choice questions. The assumption is that confused participants would sometimes side with the correct opinion and then side with the incorrect opinion at other times. Shifting between correct and incorrect opinions would lead to an overall lower rate of correct responses.

Participant responses to forced-choice questions were tracked (1 for a correct response and 0 for an incorrect response) with four mixed-effects logistic regression models (one for each trial). The unit of analysis was an individual case study, so there were 248 cases in the data set (31 learners × 8 case studies per learner). Significant models were discovered for Trial 2 ($\chi^2(3) = 31.7$, $p < .001$), Trial 3 ($\chi^2(3) = 37.0$, $p < .001$), and Trial 4 ($\chi^2(3) = 9.79$, $p < .05$) (see Table 3). As the trialogues became increasingly more specific (Trials 2–4), participants were less likely to respond correctly when they were in the experimental conditions than in the no-contradiction control condition (*true-true*).

We also compared participants' performance on the forced-choice questions following contradictions to random guessing (or chance). Since these questions adopted a two-alternative multiple-choice format, random guessing would yield a score of 0.5. This analysis involved a series of one-sample *t*-tests comparing the accuracy of participant responses on each trial to a value of 0.5.

Analyses revealed that responses varied by trial and condition when compared to chance (see Fig. 4). Performance on Trial 1 was at chance level for all conditions (p'$s > .1$). For Trial 2, performance in the *true-false* and *false-true* conditions remained at chance level (p'$s > .1$), while performance in the *true-true* condition was greater than chance ($t(30) = 9.29$, $p < .001$) and the *false-false* condition was below chance ($t(30) = -3.11$, $p = .002$). This pattern suggests that at the second trial participants were responding based on the information presented by the two agents. However, when the agents contradicted each other (*true-false* and *false-true*) participants were unsure about how to proceed.

The discussion in Trials 3 and 4 was more specific about the potential flaw in the case study, as opposed to simply noting whether there was or was not a problem with the methodology. For these two trials, all conditions significantly differed from chance performance. The *true-true* condition (Trial 3: $t(30) = 7.47$, $p < .001$; Trial 4: $t(30) = 3.72$, $p < .001$) and the *true-false* condition (Trial 3: $t(30) = 2.79$, $p = .005$; Trial 4: $t(30) = 2.04$, $p = .025$) performed above chance, whereas the *false-true* condition (Trial 3: $t(30) = -2.56$, $p = .008$) and false-false condition (Trial 3: $t(30) = -3.05$, $p = .003$; Trial 4: $t(30) = -1.31$, $p = .101$) performed below chance. There was one exception to this pattern. For Trial 4, performance in the *false-true* condition was at chance level ($t(30) = -.571$, $p = .286$). Overall, these patterns suggest that when the two agents agreed, participant responses were guided by the opinions of the agents, regardless of opinion accuracy or question specificity. However, when the agents contradicted each other, the nature of the contradiction (i.e., who takes the correct vs. incorrect position) and question specificity impacted answer correctness.

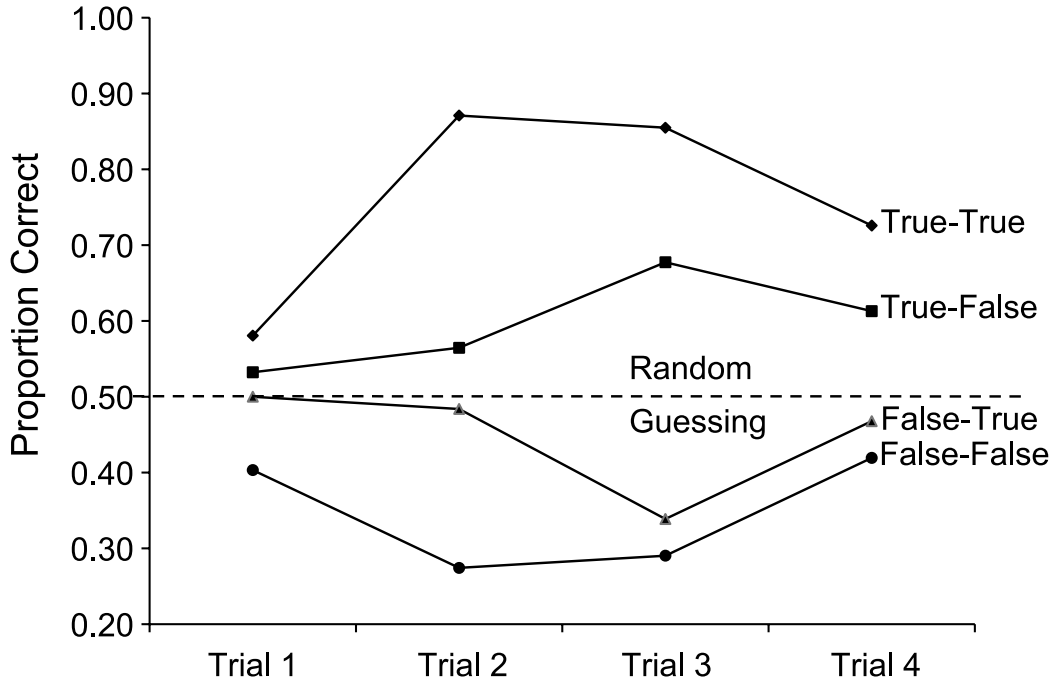

Fig. 4. Performance on forced-choice questions after contradictions in each trial.

Generally, these results seem to suggest that performance on forced-choice questions after contradictions is a potentially viable method to track confusion. In addition, the magnitude of confusion seemed to have been dependent upon several factors, including the point in the trialogue discussion and the source and severity of the contradictions.

Response shifting behavior

The decreased performance on forced-choice questions was hypothesized to occur because when participants were presented with contradictions they were switching, or shifting, between correct and incorrect responses. In the previous analyses we confirmed that when participants were in the experimental conditions they were less likely to respond correctly than the no-contradiction control. However, decreased performance could also occur if participants were simply selecting the incorrect response in each trial. To address this issue we looked at the likelihood that participants shifted their opinions during the trialogues. We hypothesized that participants would exhibit more shifts in response accuracy across trials when they were in the experimental conditions than in the control condition. A shift was defined as changing from one opinion (correct or incorrect) to the opposing opinion in successive trials (e.g., Trial 1 to Trial 2).

We investigated participants' overall shift rates (proportion of shift occurrence across trials). A mixed-effects linear regression model with shift rate as the dependent variable was significant, $F(3, 244) = 2.95$, $Mse = .218$, $p = .017$. When participants were in the *true-false* ($M = .371$, $SD = .196$) and *false-true* conditions ($M = .323$, $SD = .247$), they were significantly more likely to shift compared to the *true-true* condition ($M = .242$, $SD = .187$). The *false-false* condition ($M = .263$, $SD = .218$), however, did not differ on overall shift rate compared to the *true-true* condition.

These findings suggest that shift rates differed based on the type of contradiction. When the agents contradicted each other, the participant shifted between agreeing with each agent's opinion, presumably because

Table 4
Proportion of correct responses on the posttest

	Proportion Correct				Coefficient and odds ratio (exp(B))					
Effect	*Tr-Tr*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*	*Tr-Fl*		*Fl-Tr*		*Fl-Fl*	
Main Effect										
Condition	.419	.414	.407	.419	−.004	(.996)	−.006	(.994)	.005	(1.01)
Confusion × Condition										
Low	.460	.385	.386	.453	−.072	(.931)	−.063	(.939)	.000	(1.00)
High	.312	.474	.449	.350	**.167**	(1.18)	.129	(1.14)	.025	(1.03)

Notes. Tr: True; Fl: False; Tr-Tr was the reference group for each model, hence, coefficients for this condition are not shown in the table. Bolded cells refer to significant effects at $p < .05$.

they were confused over which was the correct opinion. When in the *false-false* condition, participants' performance on the forced-choice questions decreased but they did not shift between the correct (ground truth) and incorrect opinions (agents). In this case, participants performed poorly because they agreed with the two agents. Thus, it seems possible that the pattern of responses displayed by participants in the *true-false* and *false-true* conditions may be a more meaningful indicator of confusion than the *false-false* condition, where participants appear to simply select the majority opinion, which happens to be incorrect.

Performance on the posttest

The posttest consisted of 24 multiple-choice items, with three questions for each case study. A preliminary analysis used three mixed-effects logistic regressions to investigate condition differences on the three question types (definition, function, and example) on the posttest. None of the three models were significant, so the subsequent analyses focused on a *posttest score* that did not discriminate between the three question types. The unit of analysis was an individual case study, resulting in 248 cases in the data set. We hypothesized that confusion would have a moderating effect on learning outcomes. We predicted that the experimental conditions would perform better on the posttest than the control condition. This is because we also predicted that more confusion would be reported in the experimental conditions than the control condition.

First, a mixed-effects linear regression model with posttest score as the dependent variable and contradictory information *condition* as the fixed effect did not yield any significant differences ($p = 1.00$) (see Table 4). This analysis was limited, however, because it did not distinguish between cases in which the manipulation successfully induced confusion from cases where the manipulation was not successful. To address this issue, we conducted an analysis that investigated if levels of confusion moderated the effect of condition on performance. The analysis proceeded by dividing the 248 cases into low vs. high confusion cases based on a median split of participants' self-reported confusion for each case study. There were 158 low-confusion cases and 76 high-confusion cases. Next, a mixed-effects model with condition, confusion (low vs. high), and the condition × confusion interaction term as the fixed effects was conducted. Main effects for condition ($p = 1.00$) and confusion ($p = .802$) were not significant, but the interaction term was significant, $F(3, 240) = 2.56$, $Mse = .176$, $p = .028$.

The interaction was probed by regressing posttest scores for the low- and high-confusion cases separately. An analysis of the coefficients for both models revealed differential learning effects as a function of confusion. For the low-confusion cases, the performance of all three experimental conditions was equivalent

to the no-contradiction control. For the high-confusion cases, participants had significantly higher posttest scores when in the *true-false* condition compared to when they were in the *true-true* condition. A similar pattern emerged for the *false-true* condition; however, this effect only approached significance ($p = .091$). It appears that learners only benefited from the contradictions when confusion was *successfully* induced.

GENERAL DISCUSSION

Research on emotions during learning has targeted a set of affective states that frequently occur during learning and impact learning outcomes (Arroyo et al., 2009; Burleson & Picard, 2007; Chaffar et al., 2009; Conati & Maclaren, 2009; Craig et al., 2004; D'Mello et al., 2009; Forbes-Riley & Litman, 2011; Graesser et al., 2007; Robison et al., 2009). However, the question of how to coordinate affective and cognitive processes to increase learning still remains. One approach we have been investigating involves inducing particular affective states and then helping learners regulate these states in a manner that improves learning. Our initial focus is on confusion induction during learning. We were able to successfully induce confusion in learners through the presentation of contradictory information. Both self-reports of confusion and learner responses to forced-choice questions showed that contradictions were generally successful at inducing confusion in a context sensitive fashion. However, our findings do suggest that the ability to induce confusion is influenced by who is presenting the erroneous information (tutor vs. student agent). Furthermore, learner responses to key questions and shifting behaviors may serve as a more effective and unbiased method to track confusion compared to self-reports.

We did not expect impressive learning gains in the present study because confusion was only induced and not appropriately scaffolded. Confusion induction is likely to cause learners to stop, reflect, and deliberate, all of which are beneficial for learning. However, this depends on whether learners realize they have reached an impasse and launch effortful deliberation and problem solving activities to resolve their confusion. Confusion resolution could only occur in the present study if learners resolved confusion on their own because the learning environment provided no explicit support for confusion resolution. Hence, there are likely to be instances in which confusion was successfully induced but learners did not show improvements in learning because they were unable to resolve their confusion. In these instances learners may have benefited from some type of scaffolding, such as the sub-dialogues used in the uncertainty-adaptive version of ITSPOKE (Forbes-Riley & Litman, 2011). Nevertheless, there were modest improvements in learning, but only when confusion was successfully induced. This finding is consistent with impasse-driven theories of learning in that learners must be aware of their confusion in order to engage in the cognitive activities necessary for confusion resolution to occur (VanLehn et al., 2003).

Limitations

There are three important limitations of this research. First, critics might object to the manipulation on the grounds that we are intentionally providing misleading information and contradictions to learners and this is not in their best interest. We acknowledge this and similar reactions to the manipulation, but we feel that they are less of a concern in the present research for the following reasons: (a) any misleading information presented was corrected at the end of each trialogue, (b) there were no negative learning effects that could be attributed to the contradictions, (c) all research protocols were approved by the appropriate IRB board, (d) learners were consenting research participants instead of actual students, and (e) learners were fully debriefed at the end of the experiment. In fact, with respect to the second point, there was evidence that learners who were successfully confused by the contradictions learned more than those who were not confused.

The second limitation with the present study pertains to the lack of sensitivity of the self-report measures. Self-reports are a viable method for measuring affective experience during learning. However, there are many factors that contribute to the validity of affect self-reports that are beyond the control of the experimenters. For example, some learners may not report that they are confused due to social pressures when it comes to reporting negative affective states (Tourangeau & Yan, 2007).

When self-report measures are used to track affective states, it is difficult to separate learners' actual experience of an affective state from their willingness to report experiencing the affective state (Rasinski, Visser, Zagatsky, & Rickett, 2005). Robison, McQuiggan, and Lester (2008), for example, identified personality profiles that were related to different self-reported affective experiences during learning. The personality profiles were based on gender, goal orientation (Elliot & McGregor, 2001), and the five-factor model of personality (McCrae & Costa, 2003). They found that learners who were less agreeable, less conscientious, and had greater emotional stability reported more confusion than their counterparts. However, was it the case that these learners experienced more confusion or were simply more willing to report experiences of confusion? Due to this potential problem with self-report measures, we believe that coupling self-reports with more objective measures (e.g., facial expression, performance during the trialogue, etc.) is the most defensible position.

The third limitation relates to the generalizability of the present findings. The present study evaluated contradictory information as a method of confusion induction on a small sample ($N = 31$). The present findings should be interpreted with a modicum of caution due to this small sample size. A replication of the present findings is warranted to determine the effectiveness of this method of confusion induction and to better understand the causal relationship between confusion and learning.

Future directions

Since we have had some success at inducing confusion, the next step is to implement interventions that will take advantage of these learning opportunities. A learning environment that detects learner confusion has a variety of paths to pursue. Empathic agents are one method of intervention that has been investigated in response to a variety of learner affective states (e.g., confused, bored, curious, etc.) (Robison et al., 2009). When learners self-reported their current affective state, the agent could either adopt a parallel or reactive empathetic response. In a parallel empathetic response the agent mirrors the learner's affective state, whereas the agent displays the desired affective state in the reactive empathetic response. For example, if the desired affective state is curious but the learner is currently frustrated, the agent could either be frustrated as well (parallel) or the agent could be curious (reactive). Empathic agents were able to influence transitions between affective states during learning; however, this method was not successful across all affective states experienced by learners.

Another approach is to use an intervention method that specifically targets confusion. In one method of confusion regulation, the learning environment could keep the learner confused (i.e., in a state of cognitive disequilibrium) and leave it to the learner to actively deliberate and reflect on how to restore equilibrium. This view is consistent with a Piagetian theory (1952) that stipulates that learners need to experience cognitive conflict for a sufficient amount of time before they adequately deliberate and reflect via self-regulation. If so, the learning environment should give indirect hints and generic pumps to get the learner to do the talking when floundering. VanLehn et al. (2003) proposed a similar strategy that places the responsibility for impasse resolution on the learner, with minimal input from the tutor. This strategy entails (a) prompting the learner to reason and arrive at a solution, (b) prompting the learner to explain their solution, and (c) providing the solution with an explanation only if the learner fails to arrive at an answer.

Alternatively, Vygotskian theory (1978) suggests that it is not productive to have low ability learners spend a long time experiencing negative affect in the face of failure. If so, the learning environment should

give more direct hints and explanations. This type of intervention may be more suited for not only low ability learners, but also learners that are experiencing persistent confusion. More information or more adaptive scaffolding may be necessary for learners who are unable to resolve their confusion on their own (Forbes-Riley & Litman, 2011; Lehman et al., 2008). The uncertainty-adaptive version of ITSPOKE described earlier utilizes this type of intervention (Forbes-Riley & Litman, 2011). Based on the severity of the detected impasse, ITSPOKE engages the learner in different types of sub-dialogues that (a) draw the learner's attention to the impasse, (b) provide additional hints and questions to the learner, and (c) provide the learner with additional information that is necessary to resolve the impasse. This response to learner impasses was derived from human-human tutoring sessions (Forbes-Riley & Litman, 2007) and is similar to expert human tutor responses to confusion found in a different corpus of human-human tutoring sessions (Lehman et al., 2008).

Unfortunately, it is unlikely that there will be a one-size-fits-all intervention for technology-enhanced affect regulation. First, it may be that different affective states (e.g., boredom, confusion) require different types of intervention (Robison et al., 2009) and in the case of some affective states (e.g., flow) it may be best to not intervene at all. However, even when a single affective state is the focus (i.e., confusion), there are still likely to be learner characteristics that impact how effective an intervention is for an individual learner. Another factor that must be explored in future research is the impact of confusion induction and regulation on learners' self-efficacy and desire to interact with the learning environment. It may be the case that for some learners, confusion induction will pose too great a challenge and they will feel as though they are unable to complete the task successfully, even if scaffolds are present to aid confusion resolution. Thus, it is not only important for future research to determine which confusion interventions are effective, but also for whom each intervention is most effective.

ACKNOWLEDGMENTS

We thank our research colleagues in the Emotive Computing Group at the University of Memphis (http://emotion.autotutor.org). Special thanks to Rebekah Combs, Rosaire Daigle, Nia Dowell, Kimberly Vogt, and Lydia Perkins for data collection. This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1235958) and Institute of Education Sciences (IES R305B070349). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF or IES.

APPENDIX A

Example Test Items

Construct Validity Items

A measure that has construct validity ——— **(definition)**

a. provides a consistent measure over time.
b. actually measures what it is supposed to measure. **(correct)**
c. resembles "real world" activities.
d. provides correspondence between the independent and dependent measures.

Why is it important to have a measure that is directly linked to the construct? **(function)**

a. To show what is true or false.
b. To make sure that the results represent what they should. **(correct)**
c. So that the experiment is well grounded.
d. To show that a construct works under certain situations and settings.

A diet supplement is on the market that supposedly improves one's memory. The manufacturer relies on testimony from costumers that they feel like their memories are improved. The FDA has told the manufacturers that they must stop selling the drug unless strong evidence is provided that the drug actually improves memory for events from one's childhood. They decide to run a study to show that the drug improves memory for word meanings. Which of these possible tests would best match the construct of interest in this study? **(example)**

a. A test that has been shown to be related to recognition and recall performance **(correct)**
b. A test that measures one's ability to keep information in short-term memory
c. A test that has items that look like those seen on tests in school
d. A test on which people tend to perform consistently over time

Experimenter Bias Items

Experimenter bias refers to ——— **(definition)**

a. when the experimenter purposely misinterprets data in order to gain financial benefits.
b. when the experimenter fails to correctly operationally define the dependent variable.
c. when the experimenter intentionally uses random assignment to groups in an experiment.
d. when the experimenter affects how participants respond in the experiment. **(correct)**

Which of the following best reflects a primary concern if there is experimenter bias present in a study? **(function)**

a. The results of a study will not be accurate. **(correct)**
b. The results of a study will not be reliable.
c. The dependent variable will not be operationally defined.
d. The independent variable will not be operationally defined.

Which study most likely has a problem with experimenter bias? **(example)**

a. A soda company hires a university researcher to conduct a taste test study.
b. A soda company has their own research team conduct a taste test study. **(correct)**
c. A soda company pays actors to portray subjects in a fake taste test study.
d. A soda company conducts a taste test study that does not replicate results from prior studies.

Random Assignment Items

Random assignment refers to ——— **(definition)**

a. a procedure for assigning participants to different levels of the dependent variable to insure a normal distribution.
b. a procedure for assigning participants to ONLY the experimental condition to ensure that they are not different form one another.
c. a procedure for assigning participants to ONLY the control condition to ensure that they are not different from one another.

d. a procedure for assigning participants to the experimental and control group so they have an equal chance to be in each group. **(correct)**

Random assignment is important because ——— **(function)**

a. it ensures that the experimental and control groups are different so that the manipulation will most likely work.
b. it ensures that the experimental and control groups are similar so that the results are due to the manipulation. **(correct)**
c. it ensures that the experimental and control groups are different so that the dependent measure will differentiate between them.
d. it insures that the experimental and control groups are the same so that it is possible to manipulate the independent variable.

Which study most likely did NOT involve random assignment to the experimental or control groups. **(example)**

a. A medical researcher flips a coin to determine if participants will receive the experimental drug or a placebo.
b. A psychologist alternates experimental and control survey packets on desks in a classroom where the experiment will be run.
c. A biologist blindly assigns seeds to sample soils in order to study the impact of sample type on plant growth.
d. A psychologist assigns experimental conditions to take place in the morning and control conditions in the evening. **(correct)**

REFERENCES

Arroyo, I., Woolf, B., Cooper, D., Burleson, W., Muldner, K. & Christopherson, R. (2009). Emotion sensors go to school. In V. Dimitrova, R. Mizoguchi, B. Du Boulay & A. C. Graesser (Eds.), *Proceedings of 14th International Conference on Artificial Intelligence in Education*, Amsterdam: IOS Press. (pp. 17–24).

Barth, C. M., Funke, J. (2010). Negative affective environments improve complex solving performance. *Cognition and Emotion, 24*(7), 1259–1268.

Bates, D. M. & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. Retrieved from http://CRAN.R-project.org/package=lme4.

Bhatt, K., Evens, M. & Argamon, S. (2004). Hedged responses and expressions of affect in human/human and human/computer tutorial interactions. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of Cognitive Science Society*, Mahwah, NJ: Lawrence Erlbaum Associates, Inc. (pp. 114–119).

Bjork, R. A. & Linn, M. C. (2006). The science of learning and the learning of science: Introducing desirable difficulties. *American Psychological Society Observer*, *19*, 3.

Brown, J. & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, *4*, 379–426.

Burleson, W. & Picard, R. (2007). Evidence for gender specific approaches to the development of emotionally intelligent learning companions. *IEEE Intelligent Systems, 22*(4), 62–69.

Calvo, R. & D'Mello, S. K. (Eds.). (2011). *New perspectives on affect and learning technologies*. New York, NY: Springer.

Carroll, J. & Kay, D. (1988). Prompting, feedback and error correction in the design of a scenario machine. *International Journal of Man-Machine Studies*, *28*(1), 11–27.

Chaffar, S., Derbali, L. & Frasson, C. (2009). Inducing positive emotional state in intelligent tutoring systems. In V. Dimitrova, R. Mizoguchi, B. Du Boulay & A. Graesser (Eds.), *Proceedings of 14th International Conference on Artificial Intelligence in Education*, Amsterdam: IOS Press. (pp. 716–718).

Chan, C., Burtis, J. & Bereiter, C. (1997). Knowledge building as a mediator of conflict in conceptual change. *Cognition and Instruction*, *15*, 1–40.

Chinn, C. A. & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science education. *Review of Educational Research*, *63*(1), 1–49.

Conati, C. & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, *19*(3), 267–303.

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., Craig, S., Fike, K. & Graesser, A. (2009). Responding to learners' cognitive-affective states with supportive and shakeup dialogues. In J. Jacko (Ed.), *Human-computer interaction. Ambient, ubiquitous and intelligent interaction* (pp. 595–604). Berlin/Heidelberg: Springer.

D'Mello, S. K., Craig, S. D., Sullins, J. & Graesser, A. C. (2006). Predicting affective states through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education*, *16*, 3–28.

D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B. & Graesser, A. (2008). Automatic detection of learners' affect from conversational cues. *User Modeling and User-Adapted Interaction*, *18*(1-2), 45–80.

D'Mello, S. & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognition & Emotion*, *25*(7), 1299–1308.

D'Mello, S. & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*, 145–157.

D'Mello, S. & Graesser, A. (in press). Confusion. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *Handbook of emotions and education*. New York, NY: Taylor & Francis.

D'Mello, S. & Graesser, A. C. (in review). Inducing and tracking confusion and cognitive disequilibrium with breakdown scenarios.

D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., et al. (2010). A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In J. Kay & V. Aleven (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 245–254). Berlin / Heidelberg: Springer.

Elliot, A. & McGregor, H. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology*, *80*(3), 501–519.

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row Peterson.

Forbes-Riley, K. & Litman, D. (2007). Investigating human tutor responses to student uncertainty for adaptive system development. In A. Paiva, R. Prada & R. W. Picard (Eds.), *Proceedings of International Conference on Affective Computing and Intelligent Interaction* (pp. 678–689). Berlin, Heidelberg: Springer.

Forbes-Riley, K. & Litman, D. (2011). Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech and Language*, *25*(1), 105–126.

Graesser, A. C., Chipman, P., King, B., McDaniel, B. & D'Mello, S. (2007). Emotions and learning with AutoTutor. In R. Luckin, K. Koedinger & J. Greer (Eds.), *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, Amsterdam: IOS Press. (pp. 569–571).

Graesser, A., 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 breakdown. *Memory & Cognition*, *33*(7), 1235–1247.

Graesser, A. C., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S. & Gholson, B. (2006). Detection of emotions during learning with AutoTutor. In R. Sun (Ed.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 285–290). Mahwah, NJ: Lawrence Erlbaum & Associates.

Halpern, D. F. (2003). *Thought and knowledge: An introduction to critical thinking* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Lehman, B., Matthews, M., D'Mello, S. & Person, N. (2008). What are you feeling? Investigating student affective states during expert human tutoring sessions. In B. Woolf, E. Aimeur, R. Nkambou & S. Lajoie (Eds.), *Proceedings of 9th International Conference on Intelligent Tutoring Systems* (pp. 50–59). Berlin, Heidelberg: Springer-Verlag.

Litman, D. & Forbes-Riley, K. (2004). Annotating student emotional states in spoken tutoring dialogues. In M. Strube & C. Sidner (Eds.), *Proceedings of SIGdial Workshop on Discourse and Dialogue* (pp. 144–153). Cambridge, MA: Association for Computational Linguistics.

McCrae, R. & Costa, P. (Eds.) (2003). *Personality in adulthood: A five-factor theory perspective* (2nd ed.). New York, NY: Guilford Press.

Pekrun, R. & Stephens, E. J. (2012). Academic emotions. In K. R. Harris, S. Graham, T. Urdan, S. Graham, J. M. Royer & M. Zeidner (Ed.), *APA educational psychology handbook, Vol. 2* (pp. 3–31). Washington, DC: American Psychological Association.

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

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer Verlag.

Pon-Barry, H., Schultz, K., Bratt, E. O., Clark, B. & Peters, S. (2006). Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education*, *16*, 171–194.

Porayska-Pomsta, K., Mavrikis, M. & Pain, H. (2008). Diagnosing and acting on student affect: The tutor's perspective. *User Modeling and User-Adapted Interaction*, *18*, 125–173.

Rasinski, K. A., Visser, P. S., Zagatsky, M. & Rickett, E. M. (2005). Using implicit goal priming to improve the quality of self-report data. *Journal of Experimental Social Psychology, 41*(3), 321–327.

Robison, J., McQuiggan, S. & Lester, J. (2008). Differential affective experiences in narrative-centered learning environments. In *Proceedings of the Workshop on Emotional and Cognitive Issues in ITS in conjunction with the 9th International Conference on Intelligent Tutoring Systems*. Berlin, Heidelberg: Springer-Verlag.

Robison, J., McQuiggan, S. & Lester, J. (2009). Evaluating the consequences of affective feedback in intelligent tutoring systems. In C. Muhl, D. Heylen, A. Nijholt (Eds.), *Proceedings of International Conference on Affective Computing & Intelligent Interaction* (pp. 37–42). Los Alamitos, CA: IEEE Computer Society Press.

Rodrigo, M. M. T. & Baker, R. S. J. d. (2011a). Comparing the incidence and persistence of learners' affect during interactions with different educational software packages. In R. Calvo & S. D'Mello (Eds.), *New perspective on affect and learning technologies* (pp. 183–200). New York, NY: Springer.

Rodrigo, M. M. T. & Baker, R. S. J. d. (2011b). Comparing learners' affect while using an intelligent tutor and an educational game. *Research and Practice in Technology Enhanced Learning*, *6*(1), 43–66.

Roth, K. J., Druker, S. L., Garnier, H. E., Lemmens, M., Chen, C. & Kawanaka, T. (2006). *Teaching science in five countries: Results From the TIMSS 1999 video study* (NCES 2006-011). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Silvia, P. J. (2010). Confusion and interest: The role of knowledge emotions in aesthetic experience. *Psychology of Aesthetics Creativity and the Arts, 4*, 75–80.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T. & Baggett, W. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, *21*(3), 209–249.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.