Confusion Can Be Beneficial for Learning

Sidney D'Mello[1], Blair Lehman[2], Reinhard Pekrun[3], and Art Graesser[2]

[1]University of Notre Dame

[2]University of Memphis

[3]University of Munich

Author Note

This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1108845) and Institute of Education Sciences (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.

Corresponding author:
Sidney D'Mello
384 Fitzpatrick, University of Notre Dame, Notre Dame, IN 56556, USA
Phone: 901-378-0531, Fax: 574-631-8883, Email: sdmello@nd.edu

Abstract

We tested key predictions of a theoretical model positing that confusion, which accompanies a state of cognitive disequilibrium that is triggered by contradictions, conflicts, anomalies, erroneous information, and other discrepant events, can be beneficial to learning if appropriately induced, regulated, and resolved. Hypotheses of the model were tested in two experiments where learners engaged in trialogues on scientific reasoning concepts in a simulated collaborative learning session with animated agents playing the role of a tutor and a peer student. Confusion was experimentally induced via a contradictory information manipulation involving the animated agents expressing incorrect and/or contradictory opinions and asking the (human) learners to decide which opinion had more scientific merit. The results indicated that self-reports of confusion were largely insensitive to the manipulations. However, confusion was manifested by more objective measures that inferred confusion on the basis of learners' responses immediately following contradictions. Furthermore, whereas the contradictions had no effect on learning when learners were not confused by the manipulations, performance on multiple-choice posttests and on transfer tests was substantially higher when the contradictions were successful in confusing learners. Theoretical and applied implications are discussed.

Keywords: confusion, emotions, learning, scientific reasoning, impasses, cognitive disequilibrium

Confusion Can Be Beneficial for Learning

It is often assumed that confidence and certainty are preferred over uncertainty and confusion during learning. It is also frequently assumed that the role of a human mentor or intelligent educational technology is to dynamically tailor the instruction to match the knowledge and skills of the learner, so that states of uncertainty and confusion can be minimized. Furthermore, if and when a learner does get confused, the common instructional strategy is to quickly identify the source of the confusion and provide explanations and other scaffolds to alleviate the confusion. Simply put, common wisdom holds that confusion should be avoided during learning and rapidly resolved if and when it arises.

It might be the case, however, that these widely held assumptions are too simplistic and somewhat inaccurate. Perhaps confusion can and should be avoided for simple learning tasks such as memorizing content to be reproduced later, repeated practice of learned skills, and rote-application of learned procedures to new situations that differ on minor surface-level features but are structurally similar to the training situations. However, it is unlikely that confusion can be avoided for more complex learning tasks, such as comprehending difficult texts, generating cohesive arguments, solving challenging problems, and modeling a complex system. Complex learning tasks require learners to generate inferences, answer causal questions, diagnose and solve problems, make conceptual comparisons, generate coherent explanations, and demonstrate application and transfer of acquired knowledge (Graesser, Ozuru, & Sullins, 2010). This form of deep learning can be contrasted with shallow learning activities (memorizing key phrases and facts) and simple forms of procedural learning. Confusion is expected to be more the norm than the exception during complex learning tasks. Moreover, on these tasks, confusion is likely to promote learning at deeper levels of comprehension under appropriate conditions, as discussed in more detail below.

There is considerable empirical evidence to support the claim that confusion is prevalent during complex learning. Most compelling is the fact that confusion was the second (out of 15) most frequent emotion in a recent meta-analysis of 21 studies from the literature that systematically monitored the emotions of 1430 learners over the course of 1058 hours of interactions with a range of learning technologies, including intelligent tutoring systems, serious games, simulation environments, and computer interfaces for problem solving, reading comprehension, and argumentative writing (D'Mello, in review). The proportional occurrence of confusion with respect to the total number of emotion reports in individual studies ranged from 3% to 50%, with a normalized (across studies) mean of 15%. In addition to its prevalence during human-computer learning sessions, confusion has also been found to be prevalent during human-human tutoring sessions. For example, Lehman and colleagues (Lehman, D'Mello, & Person, 2010; 2008) analyzed the emotions that learners experienced during 50 hours of interactions with expert human tutors and found that confusion was the most frequent emotion. The prevalence of confusion during complex learning activities motivated the present focus on this emotion.

1.1. Theoretical Framework and Previous Research

The theoretical status of confusion in the affective sciences is quite mixed. Confusion has been considered to be a bona fide emotion (Rozin & Cohen, 2003), a knowledge emotion (Silvia, 2010), an epistemic emotion (Pekrun & Stephens, 2012), an affective state but not an emotion (Keltner & Shiota, 2003), and a mere cognitive state (Clore & Huntsinger, 2007). D’Mello (in press) argues that confusion meets several of the important criteria to be considered an emotion. In the present study we consider confusion to be an epistemic or a knowledge emotion (Pekrun & Stephens, 2012; Silvia, 2010) because it arises out of information-oriented appraisals of the extent to which incoming information aligns with existing knowledge structures and whether there are inconsistencies and other discrepancies in the information stream.

Mandler's interruption (discrepancy) theory (Mandler, 1984, 1990) provides a useful sketch of how confusion arises from information and goal-oriented appraisals. According to this theory, individuals are continually assimilating new information into existing knowledge structures (e.g., existing schemas or mental models) when they are engaged in a complex learning task. When discrepant information is detected (e.g., a conflict with prior knowledge), attention shifts to discrepant information, the autonomic nervous systems increases in arousal, and the individual experiences a variety of possible emotions, depending on the context, the amount of change, and whether important goals are blocked (Stein & Levine, 1991).

Surprise is likely to be the first reaction when there is a discrepancy between prior expectations and the new information. Confusion is hypothesized to occur when there is an ongoing mismatch between incoming information and prior knowledge that cannot be resolved right away, when new information cannot be integrated into existing mental models, or when information processing is interrupted by inconsistencies in the information stream. These are all conditions that instigate cognitive disequilibrium (incongruity, dissonance, conflict). Discrepant events that induce cognitive disequilibrium can include obstacles to goals, interruptions of organized action sequences, impasses, contradictions, anomalous events, dissonance, unexpected feedback, exposure of misconceptions, and general deviations from norms and expectations (Chinn & Brewer, 1993; Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005; Graesser & Olde, 2003; Mandler, 1999; Piaget, 1952; Siegler & Jenkins, 1989; Stein, Hernandez, & Trabasso, 2008; Stein & Levine, 1991; VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003). These events frequently occur over the course of performing a difficult learning task, so it is no surprise that confusion is prevalent during complex learning activities.

In addition to the incidence of confusion during complex learning, previous research has indicated that confusion is positively related to learning outcomes. Craig et al. (2004) conducted an online observational study in which the affective states (frustration, boredom, engagement/flow, confusion, eureka) of 34 learners were coded by observers every five minutes during interactions with AutoTutor, an intelligent tutoring system (ITS) with conversational dialogues (Graesser, Chipman, Haynes, & Olney, 2005). Learning gains were measured via pre and posttests that were administered before and after the learning session, respectively. The results indicated that learning gains were positively correlated with confusion and engagement/flow, negatively correlated with boredom, and were uncorrelated with the other emotions. Importantly, when learning gains were regressed on the incidence of the individual emotions, confusion was the only emotion that significantly predicted learning. This initial finding of a positive correlation between confusion and learning has subsequently been replicated in follow-up studies with AutoTutor that used different versions of the tutor, different input modalities (i.e., spoken vs. typed responses), and different methods to monitor emotions (D’Mello & Graesser, 2011; Graesser, Chipman, King, McDaniel, & D'Mello, 2007).

The positive relationship between confusion and learning is consistent with theories that highlight the merits of impasses and activating emotions during learning (Brown & VanLehn, 1980; VanLehn et al., 2003). Impasse-driven theories of learning posit that impasses (and the associated state of confusion) are beneficial to learning because they provide learning opportunities. That is, once an impasse is detected and confusion is experienced, the individual needs to engage in effortful cognitive activities in order to resolve their confusion. Confusion resolution requires the individual to stop, think, engage in careful deliberation, problem solve, and revise their existing mental models. These activities involve desirable difficulties (Bjork & Bjork, 2011; Bjork & Linn, 2006), which inspire greater depth of processing during training, more durable memory representations, and more successful retrieval (Craik & Lockhart, 1972; Craik & Tulving, 1972). Evidence for impasse-driven learning can be found in early work on skill acquisition and learning (Brown & VanLehn, 1980; Carroll & Kay, 1988; Siegler & Jenkins, 1989) and in more recent work in problem solving (D'Mello & Graesser, in review), one-on-one tutoring (Forbes-Riley & Litman, 2009, 2010; VanLehn et al., 2003), and conceptual change (Chi, 2008; Dole & Sinatra, 1998; Nersessian, 2008). For example, in an analysis of approximately 125 hours of human-human tutorial dialogs, VanLehn et al. (2003) found that comprehension of physics concepts was rare when learners did not reach an impasse, irrespective of the quality of the explanations provided by tutors.

It is important to note that all instances of confusion are not alike and are not expected to have equivalent effects on learning. For example, a tutor can induce confusion by intermittently speaking in a foreign language or a learner might be uncertain about the location of his or her mathematics textbook. These instances of confusion that are peripheral to the learning activity are unlikely to have any meaningful impact on learning. It is the instances of confusion that are contextually coupled to the learning activity that are of importance. Even so, all contextualized instances of confusion are not expected to impact learning in similar ways. For example, a learner who gets confused by a difficult math problem but disengages after making a few unsuccessful attempts to solve the problem is not expected to learn anything substantial. Similarly, persistent confusion, which occurs when confusion resolution fails, is expected to accompany negligible or poor learning when compared to situations where confusion is immediately or eventually resolved (D'Mello & Graesser, 2012). In the VanLehn et al. (2003) tutoring example discussed earlier, learners acquired a physics principle in only 33 of the 62 impasse occurrences, ostensibly because their impasses were not resolved for the remaining 29 cases. Therefore, it would be worthwhile to distinguish between productive and unproductive confusion.

This distinction was recently tested in two experiments that induced confusion while participants performed a device comprehension task (understanding how devices such as toasters and doorbells work from technical illustrated texts) (D'Mello & Graesser, in review). The manipulation consisted of presenting participants with descriptions of device malfunctions (e.g., “When a person rang the bell there was a short *ding* and then no sound was heard.”) and asking them to diagnose the problem. A second-by-second analysis of the dynamics of confusion yielded two characteristic trajectories that successfully distinguished those individuals who partially resolved their confusion (confusion initially peaked and then dampened) over the course of diagnosing the breakdowns from those who remained confused (confusion continued to increase). As predicted, individuals who partially resolved their confusion performed significantly better on a subsequent device comprehension test than individuals who remained confused.

To summarize, there is considerable theoretical justification and some empirical support to suggest that confusion plays an important role during complex learning activities. The theoretical model we adopt posits that one important form of deep learning occurs when there is a discrepancy in the information stream and the discrepancy is identified and corrected. Otherwise, the person already has mastered the task and by definition there is no learning, at least from a perspective of conceptual change (Chi, 2008; Chinn & Brewer, 1993; Nersessian, 2008). A major assertion that emerges from the model is that learning environments need to substantially challenge learners in order to elicit critical thought and deep inquiry. The claim is that confusion can be beneficial to learning if appropriately regulated because it can cause individuals to process the material more deeply in order to resolve their confusion. It should be emphasized that we do not expect that confusion itself is sufficient to promote meaningful conceptual change. Quite different from this, the present claim is more modest in that confusion is expected to have a measurable positive impact on learning because it serves as a catalyst to engender deeper forms of processing. There is some correlational evidence to support this claim (Craig et al., 2004; D'Mello & Graesser, 2011; Forbes-Riley & Litman, 2009, 2010; Forbes-Riley & Litman, 2011; Graesser et al., 2007; VanLehn et al., 2003), but to our knowledge, there is no causal evidence linking confusion to learning outcomes. In essence, the causal claim that confusion can lead to enhanced learning remains untested. This is somewhat surprising because it has been more than a century since initial observations into the role of confusion (expressions) in deep thought and inquiry were reported (Darwin, 1872). The goal of the present paper is to address this gap in the literature by causally linking confusion and deep learning.

1.2. Hypotheses and Overview of Present Research

The working theoretical model posits that confusion can positively impact learning if (a) it is appropriately induced in context and (b) learners have the ability to appropriately resolve confusion, or (c) the learning environment provides sufficient scaffolds to help learners resolve the confusion when they cannot resolve it on their own. We tested this key prediction of the model in a unique learning environment consisting of learners engaging in *trialogues* on scientific reasoning concepts (e.g., construct validity, experimenter bias) with two animated pedagogical agents that simulated a tutor and a peer student. The trialogues consisted of the agents and the human learner evaluating the methodologies of scientific studies and attempting to identify flaws in the design of the studies. Confusion was induced with a *contradictory information manipulation* in which the tutor and peer student agents staged a disagreement on the quality of the study (one was correct and the other was incorrect) and eventually invited the learner to intervene. In other situations the two agents agreed on a fact that was patently incorrect. Here, the contradiction was not between the agents but between the ground truth and the erroneous opinion expressed by both agents. Both forms of contradictions were expected to trigger cognitive disequilibrium and confusion at multiple levels (Hypothesis 1: *contradictory information hypothesis*). There is disequilibrium at the *cognitive level* because the presence of the contradiction signals a discrepancy in the information stream which presumably violates the learner's expectations. This is consistent with the classical Piagetian (1952) theory of cognitive disequilibrium. However, the simulated collaborative learning environment is also expected to create disequilibrium at the *socio-cognitive level* (Mugny & Doise, 1978). This would occur when there is disagreement between the two agents, when the learner disagrees with one of the agents, or when the learner disagrees with both agents.

Confusion is thought to force the learner to reflect, deliberate, and decide which opinion had more scientific merit. The hypothesis is that learners who are confused would be more vigilant and process the material at deeper levels of comprehension than learners who are not confused. This form of deeper processing that is launched when learners are confused is expected to positively impact learning. However, this positive impact is only expected to occur if learners can effectively self-regulate their confusion or if the learning environment provides sufficient scaffolds to help learners regulate their confusion (Hypothesis 2: *facilitative confusion hypothesis*). This is again consistent with theories of cognitive disequilibrium and socio-cognitive disequilibrium when cognitive restructuring is facilitated by information transmitted over the course of the trialogues (Dimant & Bearison, 1991; Mugny & Doise, 1978).

It is also likely that prior knowledge moderates the effect of the contradictions on confusion and learning. It takes a modicum of knowledge to know what one knows and does not know (Miyake & Norman, 1979), so it might be the case that only learners with some domain-knowledge will experience confusion when confronted by the contradictions. Low-domain knowledge learners might simply fail to detect the impasse and will not experience confusion (VanLehn et al., 2003). Even in cases where a contradiction succeeds at confusing a low domain-knowledge learner, the learner might not have the necessary acumen to effectively resolve the confusion, so the contradiction is likely to have negligible effects on learning gains. In essence, prior knowledge is expected to influence the extent to which the *contradictory information* and *facilitative confusion* hypotheses are supported.

The two hypotheses were tested in two experiments. In Experiment 1, 64 participants engaged in trialogues with the animated agents and attempted to detect flaws in sample research studies. There were four opportunities for contradictions during the discussion of each research study. Experiment 2 (N = 76) had a *delayed contradiction manipulation*, where the animated agents initially agreed with each other, but eventually started to express divergent views. Experiment 2 also tested whether *reading text* would help as an intervention to alleviate confusion. Specifically, learners were first put into a state of confusion via delayed contradictions and were then asked to read a text that could potentially resolve their confusion. The tutor agent concluded each trialogue by presenting the correct information in order to mitigate any adverse effects of the contradictions and to give learners a final opportunity to inspect and revise their mental models.

Learners’ confusion levels were measured via retrospective affect judgment (Experiment 1) and online self-report (Experiment 2) protocols. Confusion was also indirectly inferred via the accuracy of learners' responses to forced-choice questions immediately following contradictory statements (both experiments). Learning was measured with multiple-choice posttests that assessed shallow knowledge of scientific reasoning concepts (both experiments) and with near and far transfer versions of a flaw detection task (Experiment 2).

2. Learning Content and Learning Environment

The learning domain for the present experiments was critical thinking and scientific reasoning. Scientific reasoning and inquiry involves conceptual skills related to designing and evaluating experiments, such as stating hypotheses, identifying dependent and independent variables, isolating potential confounds in designs, interpreting trends in data, determining if data support predictions, and understanding effect sizes (Halpern, 2003; Roth et al., 2006). The processes of evaluating a study scientifically and asking critical questions are fundamental to scientific inquiry. Hence, the learning environment attempted to teach fundamental scientific inquiry skills by presenting example case studies (including the research design, participants, methods, results, and conclusions) that were frequently flawed. Learners were instructed to evaluate the merits of the studies and point out problems. These critiques were accomplished by holding multi-turn trialogues with two embodied conversational agents and the human learner.

One agent called the *tutor agent*, or *Dr. Williams*, leads the tutorial lessons and is an expert on scientific inquiry. The second agent, *Chris*, is the *peer-agent* who simulates a peer of the human learner (i.e., the participant in the experiment). The human learners interact with both agents by holding dialogues in natural language that mimic real tutorial interactions. The tutor agent gives the human learner and peer agent descriptions of research studies and texts to read, poses diagnostic questions and situated problems, asserts her opinion, and provides explanations (at very specific points in the trialogues as discussed below).

An excerpt of the trialogues between the two agents and the human learner (Bob) is presented in Table 1. Each learning session began with a description of a sample case study. The case study in the excerpt pertained to random assignment and is flawed because participants were not randomly assigned to conditions. Note that the human is referred to as *human learner, learner*, or *participant*. Peer student or peer agent refers to the animated conversational agent who is a virtual peer of the human learner.

Learners were then asked to read the study in order to familiarize themselves with the specifics of the study before discussing its scientific merits. The discussion of each study occurred over four (Experiment 1) or five (Experiment 2) multi-turn trials. For example, dialogue turns 4 through 8 in Table 1 represent one trial. Each trial consisted of the peer, Chris, (turn 5) and the tutor, Dr. Williams, (turn 6) agents asserting their opinions on one facet of the study. The tutor agent prompted the human learner (Bob in this case) to provide his opinion (turn 7) and obtained the learner's response (turn 8). As will be described in the next section, confusion was induced by manipulating the content of the agents' utterances in each trial.

The responses of the agents were pre-scripted and a multimodal interface rendered the scripts in real time. The interface shown in Figure 1 consisted of the tutor agent (A), the peer agent (B), a description of the research case study (C), a text-transcript of the dialogue history (D), and a text-box for learners to enter and submit their responses (E). The agents delivered the content of their utterances via synthesized speech while the human learner typed his or her responses. Text-transcripts of the trialogues were stored in log files for offline analysis.

3. Experiment 1

3.1. Method

3.1.1. Participants. Participants were 63 undergraduate psychology students from a mid-south university in the U.S. There were 21 males and 42 females in the sample. Participants’ age ranged from 18 to 50 years old (*M* = 21.0, *SD* = 5.03). Forty-eight percent of participants were Caucasian, 49% were African-American, and 3% were Asian. The participants received course credit for their participation. Prior coursework in critical thinking and scientific reasoning was not required. Eighty-four percent of participants had not taken a research methods or statistics course prior to participation.

3.1.2. Design. The experiment had a within-subjects design with four conditions: *true-true*, *true-false*, *false-true,* and *false-false* that are described below. Participants completed two learning sessions in each of the four conditions with a different scientific reasoning concept in each session (8 sessions in all). The eight concepts were construct validity, control groups, correlational studies, experimenter bias, generalizability, measure quality, random assignment, and replication. Each concept had an associated case study that might or might not have been flawed. 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 were counterbalanced across participants with a Graeco-Latin Square. Flaws were also counterbalanced across case studies and conditions.

3.1.3. Manipulation. Confusion was experimentally induced via a contradictory information manipulation. This manipulation was achieved by having the tutor and peer agents stage a disagreement on a concept and then invite the human learner to intervene. The contradiction was expected to trigger conflict and force the learner to reflect, deliberate, and decide which opinion had more scientific merit. In other words, learners had to decide if they agreed with the tutor agent, the peer 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 peer agent agreed with the tutor. The *true-true* condition served as the no-contradiction control condition. In the *true-false* condition, the tutor presented a correct opinion and the peer agent disagreed by presenting an incorrect opinion. In contrast, it was the peer 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 peer agent agreed. Although there were no contradictions in this condition, both agents provided erroneous opinions that contradicted the ground truth. All incorrect information was corrected over the course of the trialogues and participants were fully debriefed at the end of the experiment.

3.1.4. Knowledge tests. Scientific reasoning knowledge was tested before and after learning sessions with two knowledge tests (pretest and posttest, respectively). The pretest consisted of one four-alternative multiple-choice (4AFC) question for each concept discussed in the learning sessions, thereby yielding eight questions. These questions were relatively shallow and targeted participants' knowledge of the definitions of the scientific reasoning concepts.

The posttests consisted of one definition question, two function questions, and two example questions for each of the eight concepts, thereby yielding 40 questions. The definition questions consisted of eight items that were different from the pretest items. The function questions targeted the utility or function of each concept, whereas the example questions involved applications of the concepts. Examples of each question type are listed in the Appendix.

3.1.5. Procedure. Participants were individually tested in 2-2.5 hour sessions. The experiment occurred over two phases: (1) knowledge assessments and learning session and (2) retrospective affect judgment protocol.

Knowledge assessments and learning session. Participants first 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. Participants completed the pretest and were then asked to read a short introduction to critically thinking about scientific studies in order to familiarize them with the concepts that would be discussed.

Participants were instructed to put on a set of headphones after completing the pretest. The agents introduced themselves, discussed their roles, discussed the importance of developing scientific reasoning skills, and described the learning activity. Participants then analyzed eight sample research studies for approximately 50 minutes. Participants completed the posttest immediately after discussing the eighth case study.

The trialogue for each case study was comprised of four multi-turn trials (see Table 1). 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 human learner for his or her opinion via a forced-choice question, (4) the learner provided his or her opinion, and (5) learners were asked to self-explain their opinion (third and fourth trials only).

The trialogues were organized so that the specificity of the discussion increased across trials. As an example, consider the trialogue in Table 1 that pertains to a study that made a causal claim but did not use random assignment. Trial 1 was quite general and required learners to provide their opinions as to whether they would change their behavior based on the results of the study (turns 1-3). Trial 2 was a bit more direct and focused on whether there was a problem with the methodology of the study (turns 5-8). Trial 3 was more specific because it addressed the issue of whether the two groups were equivalent (turns 10-13). Finally, the target concept of random assignment was directly addressed in Trial 4 (turns 14-17).

Learners were also required to provide an explanation about their response to the forced-choice questions after trials 3 and 4 because the discussion was more specific for these trials. For example, after the learner responded “don’t know” in Trial 3, the tutor agent would say: “Bob, tell me more about why you think that" (not shown in Table 1).

All contradictory and false information was corrected after Trial 4. This consisted of the agent(s) who asserted the incorrect information acknowledging that he or she was mistaken and the tutor agent providing an accurate evaluation of the case study.

Three streams of information were recorded during the learning session. First, a video of the participant’s face was recorded with the webcam that was integrated into the computer monitor. Second, a video of the participant’s computer screen was recorded with Camtasia

Studio™. The captured video of the computer screen also included an audio stream of the synthesized speech generated by the agents. Third, the text of the trialogue, including the participant’s responses, was stored in a log file.

Retrospective affect judgment protocol. Participants provided retrospective affect judgments immediately after completing the posttest. Videos of participants' face and screen were synchronized and participants made affect ratings while viewing these videos. Participants were provided with an alphabetized list of affective states (anxiety, boredom, confusion/uncertainty, curiosity, delight, engagement/flow, frustration, surprise, and neutral) with definitions. The list of emotions was motivated by previous research on student emotions during learning with technology (D’Mello & Graesser, in press; Rodrigo & Baker, 2011) as discussed in a recent meta-analysis by D'Mello (in review).

The list of emotions was explicitly defined before the participants made their judgments. Anxiety was defined as being nervous, uneasy, apprehensive, or worried. Boredom was defined as being weary or restless through lack of interest. Confusion/uncertainty 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 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 emotions, and participants’ recent memories of the interaction.

Participants selected one emotion from the list of emotions at each judgment point. The 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 emotions after both agents provided their opinions (turns 1, 6, 11, and 15 in Table 1), after the forced-choice question was posed (turns 2, 7, 12, and 16), and after learners were asked to explain their responses in Trials 3 and 4 (not shown in Table 1). Participants also reported their emotions after reading the research study, at the end of the learning session when the tutor agent stated whether the study was flawed or not flawed, and after the tutor agent explained the scientific merits of the study (not shown 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.

3.2. Results

There were three sets of dependent measures in the present analyses: (1) self-reported affect obtained via the retrospective judgment protocol, (2) learners' responses to the tutor agent's forced-choice questions for Trials 1-4, and (3) performance on the multiple-choice posttest. The primary analyses consisted of testing for condition differences on the dependent variables and testing whether induced confusion moderated the effect of the contradictions on learning outcomes. We also tested whether prior knowledge moderated the effect of condition on the dependent variables.

Due to the repeated measurements and nested structure of the data (trials nested within case studies, case studies nested within conditions), a mixed-effects modeling approach was adopted for all analyses. Mixed-effects modeling is the recommended analysis method for this type of data (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 computation.

Linear or logistic models were constructed on the basis of whether the dependent variable was continuous or binary, respectively. The random effects were: *participant* (64 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 apriori comparison of each experimental condition to the no-contradiction control, so the *true-true* condition was the reference group in all the models. The hypotheses specify the direction of the effect, so one-tailed tests were used for significance testing with an alpha level of 0.05.

3.2.1. Self-reported affect. The retrospective affect judgment procedure yielded 6546 judgments at the pre-specified points and 296 judgments that were voluntarily provided by the learners. Due to the small number of voluntary judgments, they were combined with the fixed judgments, thereby yielding a total of 6842 affect judgments. Nine mixed-effects logistic regressions that detected the presence (coded as a 1) or absence (coded as a 0) of each affective state were constructed. The unit of analysis was an individual affect judgment, so there were 6842 cases in the data set. Significant[1] models were discovered for confusion (χ^2 (3) = 11.3, p < .001), boredom (χ^2 (3) = 11.7, p = .004), and engagement/flow (χ^2 (3) = 9.55, p = .012), but not for the other six states. Importantly, these were found to be the three most frequent states that students experience during learning sessions with technology (D'Mello, in review).

[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.

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 learners self-reported significantly more *confusion* in the *true-false* condition than in the *true-true* condition. The difference between the estimates (i.e., *B* values) for this comparison was 0.329, so learners were $e^{0.329}$ or 1.4 times more likely to report confusion in the *true-false* condition compared to the *true-true* condition. In addition to confusion, learners in the *true-false* condition were significantly more likely to report higher levels of *engagement/flow* and lower levels of *boredom* compared to the *true-true* condition.

There were no significant differences in self-reported confusion and engagement/flow when the *false-true* and *false-false* experimental conditions were compared to the *true-true* control condition. However, learners in the *false-true* condition reported significantly less boredom than in the *true-true* control. In general, the results support the conclusion that the contradictions were successful in inducing confusion and promoting deeper engagement (i.e., lower boredom plus more engagement/flow), at least for the *true-false* condition.

3.2.2. Responses to forced-choice questions. Self-reports are one viable method to track confusion. However, this measure is limited by learners' sensitivity and willingness to report their confusion levels. A more subtle and objective approach is to infer confusion on the basis of learners' responses to forced-choice questions following contradictions by the animated agents. The assumption is that learners who are presumably confused by the contradiction will be *less likely* to answer these questions correctly *in the long run* because they will sometimes side with the incorrect opinion when they are unsure about how to proceed. Their oscillation back and forth between correct and incorrect responses ends up lowering the overall rate of correct decisions.

We tracked learner answers (1 for a correct response and 0 for an incorrect response) to these forced-choice questions that were systematically integrated at key points in the trialogues, with four mixed effects logistic regression models (one for each trial). The unit of analysis was an individual case study so there were 512 cases in the data set (64 learners × 8 case studies per learner). Significant (p < .05) models were discovered for all four trials (see Table 3; χ^2 (3) = 6.33, 32.5, 40.7, and 29.0 for Trials 1, 2, 3, and 4, respectively). Learners were significantly less likely to answer *correctly* after a contradiction in Trial 1 in the *false-true* and the *false-false* conditions compared to the *true-true* control condition. The results also indicated that learners provided significantly less correct answers following contradictions in the experimental conditions as the trialogues became more specific (Trials 2-4). That is, learners in all three contradiction conditions were significantly less likely to provide correct responses on Trials 2-4 compared to learners in the no-contradiction control.

The forced-choice questions adopted a two-alternative multiple-choice format, so random guessing would yield a score of 0.5. Comparisons of the accuracy of learner responses (i.e., proportions of correct responses averaged across trials, as reflected in Table 3), revealed the following pattern: (a) accuracy in both the *true-true* and *true-false* conditions was greater than guessing although *true-false* < *true-true*, (b) accuracy in the *false-true* condition was approximately equivalent to random guessing, and (c) accuracy in the *false-false* condition was lower than random guessing. Overall, the data adhered to the following pattern in terms of correctness: *true-true* > *true-false* > *false-true* > *false-false*. This pattern is intuitively plausible in the sense that answer correctness decreased as the extent of the contradictions increased. For example, the mismatch with expectations was greater when the tutor was incorrect (*false-true*) compared to when the peer was incorrect (*true-false*) because this violates expectations.

3.2.3. Performance on the posttest. The posttests consisted of 40 multiple-choice questions, with five questions for each case study. A preliminary analysis testing for condition differences on the three question types (definition, function, and example) was not significant, so the subsequent analyses focused on a *proportional performance score* that did not discriminate among the three question types. As before, the unit of analysis was an individual case study, so there were 512 cases in the data set.

A mixed-effects linear regression model with proportional performance as the dependent variable did not yield any significant differences (*p* = .246). This analysis is limited, however, because it does not separate cases when learners were confused compared to when they were not confused. This was addressed with an additional analysis that investigated if levels of confusion moderated the effect of condition on performance. The analysis proceeded by dividing the 512 cases into low vs. high confusion cases based on a median split of learners' self-reported confusion for each case study. There were 264 low-confusion cases and 248 high-confusion cases. A mixed-effects model with *condition*, *confusion* (low vs. high), and the *condition* × *confusion* interaction term did not yield a significant main effect for *condition* (*p* = .144) or *confusion* (*p* = .136), but the interaction term was significant, *F*(3, 504) = 2.76, *Mse* = .110, *p* = .021.

The interaction was probed by regressing proportional posttest scores for the low- and high-confusion cases separately. The model for the low-confusion cases was not significant (*p* = .158), which indicated that the contradictions did not affect posttest performance when they failed to confuse the learners. In contrast, a significant model was discovered for the high-confusion cases, *F*(3, 244) = 3.70, *MSe* = .123, *p* = .006. Learners who reported being confused by the contradictions in the *true-false* and *false-false* conditions had significantly higher posttest scores than learners in the *true-true* condition (see Table 4).

An alternate method to probe the interaction is to regress posttest performance on confusion independently for each condition (see Figure 2). This yielded significant models for the *true-false*, $F(1, 126) = 5.81$, $MSe = .216$, $p = .009$, and *false-false*, $F(1, 126) = 5.42$, $MSe = .236$, $p = .011$, conditions but not the *true-true* ($p = .306$) and *false-true* conditions ($p = .187$). The slopes were positive for the *true-false* ($B = .091$) and *false-false* ($B = .109$) conditions but not for the *true-true* ($B = -.018$) and *false-true* ($B = -.041$) conditions. Learners who were confused by the contradictions in the *true-false* and *false-false* conditions were *significantly more likely* to score higher on the posttest than learners who reported being less confused (see Figure 2). Levels of confusion do in fact influence learning gains.

3.2.4. Effect of prior knowledge. On average, participants answered 51% (*SD* = 21.2%) of the pretest questions correctly. We tested whether prior knowledge moderated the effect of condition on self-reported confusion, answer correctness on the four forced-choice questions, and on posttest scores by assessing whether the prior knowledge × condition interaction term significantly predicted these six dependent variables. The interaction term was not significant ($p > .05$) for any of the models.

3.3. Discussion

This experiment tested two hypotheses on the role of confusion during learning. These hypotheses posit that confusion occurs when there are contradictions and conflicts in the information stream (Hypothesis 1: *contradictory information* hypothesis) and that the induced confusion can be beneficial for learning if effectively regulated by the learner or if there are sufficient scaffolds to help learners resolve their confusion (Hypothesis 2: *facilitative confusion* hypothesis). The results partially supported both hypotheses.

Data from the retrospective affect judgment protocol partially supported the *contradictory information* hypothesis (Hypothesis 1) in that self-reported confusion in the *true-*

false condition was significantly greater than the *true-true* condition. This finding is somewhat tempered by the fact that there were no differences in confusion levels when the *false-true* and *false-false* conditions were compared to the control. The results from the objective measure that inferred confusion from the accuracy of learner responses to forced-choice questions following contradictions were considerably more promising. Learners provided less accurate responses to these questions in all three contradiction conditions compared to the no-contradiction control.

The *facilitative confusion* hypothesis (Hypothesis 2) was tested by investigating if levels of confusion moderated the effect of contradictions on learning. The significant condition × confusion interaction indicated that there was more learning in the contradiction conditions if two criteria were satisfied. First, contradictions should be successful in inducing confusion because learners only demonstrated increased learning gains in the experimental conditions when the contradictions triggered confusion. Second, the source of the contradictions matter because significant learning effects compared to the *true-true* control were only discovered for the *true-false* and *false-false* conditions.

One notable limitation with the present study was the level of confusion reported. Although self-reported confusion in one of the experimental conditions was higher than the control condition, overall confusion levels were quite low (an average of 8.88% across conditions). This was somewhat surprising because learners' responses to the forced-choice questions were considerably lower in the contradiction conditions. This suggests that learners might have been more confused than they were reporting and the lower levels of self-reported confusion might have been an artifact of the methodology.

In addition to confusion, there were also differences in boredom and engagement/flow in some of the experimental conditions. This is not a surprising finding because one of the evolutionary functions of affective states such as confusion is to trigger a sense of alertness which can enhance information pickup and action capabilities (Darwin, 1872; Izard, 2010). In the present experiment, this was less of a concern, because despite there being condition differences for boredom and engagement/flow, there were no condition differences in learning gains. More specifically, learning effects were only discovered when learners were confused by the contradictions in two of the experimental conditions.

In summary, although Experiment 1 provided some initial support in favor of the *contradictory information* and *facilitative confusion* hypotheses, it is important that these findings replicate before we can make any major claims on the status of these hypotheses. This was achieved in a second experiment (Experiment 2) that used similar methods as Experiment 1, but with two key differences. First, contradiction manipulations in Experiment 2 were delayed instead of being immediate. Specifically, while the animated agents contradicted each other on all four trials in Experiment 1, contradictions were *delayed* until the third trial in Experiment 2. That is, the two agents conveyed the same, accurate information for Trials 1 and 2, but contradicted each other for Trials 3, 4, and 5 (an additional trial). We hypothesized that confusion levels would be more intense if an initial sense of agreement between the agents and the learner was first created (Trials 1 and 2) and then abruptly violated with contradictions and disagreements (Trials 3 and 4).

The second difference was that we tested *reading text* as one intervention to alleviate confusion. This was achieved by first confusing learners with the delayed contradictions (Trials 3 and 4) and then asking them to read a short text that was relevant to the research study being discussed. We hypothesized that participants who were confused by the contradictions would process the text at deeper levels of comprehension and thereby achieve higher learning gains than participants who were not confused.

Experiment 2 also included a number of refinements to the methodology. These involved (a) expanding the knowledge assessments to include a *flaw-identification task* that required participants to identify flaws in near and far transfer versions of the case studies, (b) replacing the *offline* retrospective judgment protocol with single-item *online* assessments of confusion, and (c) reducing the number of case studies from eight to four, while simultaneously increasing the depth of the trialogues that accompanied each case study.

4. Experiment 2

4.1. Method

4.1.1. Participants. Participants were 76 undergraduate psychology students at a mid-south university in the US. There were 26 males and 50 females. Participants ranged in age from 18 to 45 years (*M* = 20.8, *SD* = 5.86). Forty-five percent of participants were Caucasian, 45.3% were African-American, and 9.30% were Asian. Participants received course credit for their participation. Prior coursework in critical thinking and scientific reasoning was not required. Ninety-three percent of participants had not taken a research methods course and 80% had not taken a statistics course.

4.1.2. Design. The experiment had a within-subjects design with four conditions (t*rue-true*, *true-false*, *false-true*, *false-false*). Participants completed one learning session on each scientific reasoning concept in each of the four conditions (4 learning sessions in all). The four concepts were control group, experimenter bias, random assignment, and replication. The four case studies associated with each concept had one flaw. Order of conditions and concepts and assignment of concept to condition was counterbalanced across participants with a Graeco-Latin Square.

4.1.3. Knowledge tests. The knowledge tests consisted of a multiple-choice pretest and posttest along with two transfer tests. The pretest was a four-foil multiple-choice test consisting of eight definition questions (same as Experiment 1). There was a 20-item multiple-choice posttest with five items for each of the four scientific reasoning concepts covered in the learning sessions. The posttest items for each concept included one definition, one function, and one example question selected from the question bank used in Experiment 1 along with two additional questions from the explanatory text that learners were asked to read while discussing each concept (discussed further below).

The *flaw-identification task* tested learners' ability to apply the scientific reasoning concepts to sample research studies. The learner was presented with a description of a previously unseen research study and was asked to identify flaw(s) in the study by selecting as many items as they wanted from a list of eight scientific reasoning concepts. This list included four concepts that could potentially be flawed (random assignment, experimenter bias, replication, control group) and four distractors (generalizability, correlational studies, construct validity, measurement quality). Learners also had the option of selecting that there was no flaw in the research study, although each study contained one or two flaws, as discussed in more detail below.

There were near and far transfer case studies. The near transfer studies differed from the studies discussed during the learning sessions only on surface features. There was one near transfer case study with one flaw for each of the four scientific reasoning concepts. The far transfer test was considerably more difficult because it required learners to detect flaws in case studies that were different in terms of both surface and structural features. There were two far transfer studies and each contained two flaws. Problems with the control group and use of random assignment constituted the flaws in one far transfer study, while experimenter bias and replication were the flaws in the second study.

4.1.4. Explanatory texts. Learners were asked to read a short explanatory text during the trialogue of each learning session. The texts contained an average of 364 words (*SD* = 41.7 words) and were adapted from the electronic textbook that accompanies the Operation ARIES! Intelligent Tutoring System (Millis et al., in press). The explanatory text provided a brief description of each scientific reasoning concept (one text for each concept). Each text consisted of the definition of the concept, the importance of the concept, and an example of how to apply the concept in research. The texts were sufficiently general and, other than focusing on the same concept (e.g., random assignment), were *not* related to the sample case study being discussed in the trialogue.

4.1.5. Procedure. Participants were individually tested over a two-hour session. After signing an informed consent and completing the pretest, participants completed four learning sessions for approximately 60 minutes. The following is a specification of the major events for each learning session:

1. The tutor agent introduced the case study and the human learner and peer agent read the case study.
2. Both agents asserted *correct* information *without contradictions* (all conditions) on one aspect of the case study and asked the learner for a response (Trial 1).
3. Same as Trial 1 but the discussion was slightly more specific (Trial 2).
4. The agents asserted correct information without contradictions (*true-true*), or incorrect information without contradictions (*false-false*), or *incorrect* and *contradictory* information (*true-false* and *false-true* conditions), and asked the learner for his or her opinion (Trial 3).
5. Same as Trial 3 but the discussion was very specific in terms of how the study was flawed (Trial 4).

6. Learners were asked to self-report their confusion at this point in time on a three-point scale (not confused, slightly confused, confused).
7. Learners were asked to read a short text pertaining to the scientific reasoning concept being highlighted in the case study.
8. Learners self-reported their confusion once more on the same scale.
9. The agents contradicted themselves once more and asked the human to intervene (Trial 5). This trial was identical to Trial 4 but occurred after the human read the explanatory text.
10. The incorrect information was eventually corrected and the tutor agent provided an accurate evaluation of the case study.

It should be noted that learners were asked to report their confusion before and after the text intervention in order to compare confusion levels at these two different points. A similar strategy was adopted for Trial 5 with respect to performance on the forced-choice questions.

4.2. Results

There were four sets of dependent measures in the present analyses: (1) self-reported confusion at two points in the learning session, (2) learners' responses to the tutor agent's forced choice (2AFC) questions for Trials 1-5, (3) performance on the multiple-choice posttest and, (4) performance on the near and far transfer tests. Similar to Experiment 1, the data were analyzed at the level of individual case studies with mixed effects models that included participant, case study, and order of case study as the random effects. There were 304 cases across the 76 participants (76 participants × 4 case studies).

4.2.1. Self-reported confusion. Confusion was self-reported before (pre-reading confusion) and after (post-reading confusion) reading the explanatory test. Before reading, 21 learners reported that they were *confused*, 45 reported being *slightly confused,* and 238 participants reported *not being confused*. To eliminate some of the skewness in the distribution, the *slightly confused* cases were grouped with the *confused* cases, thereby increasing the number of pre-reading confusion cases to 76. A mixed-effects logistic regression model with condition as the fixed effect and pre-reading confusion as a binary dependent variable was not significant (p = .494). There were 17 pre-reading confusion cases in the *true-true* condition and 16 or 17 confusion cases in each of the three experimental conditions.

The post-reading confusion data consisted of only 4 *confused* cases and 10 *slightly confused* cases, with the remainder of the 290 cases falling into the *not confused* category. Therefore, we did not systematically test for condition differences in post-reading confusion and do not consider this measure in subsequent analyses. The small number of post-reading confusion cases also made comparisons of pre-post confusion untenable.

4.2.2. Responses to forced-choice questions. Similar to Experiment 1, confusion was inferred on the basis of incorrect responses to the forced-choice questions. We constructed five mixed-effects logistic regression models (one for each trial) that predicted correct (coded as 1) or incorrect (coded as 0) responses on the basis of condition (fixed effect). Proportions of correct responses by condition and model coefficients are presented in Table 5.

Condition was not a significant predictor of answer correctness for Trial 1 (p = .403). This is what we expected because there were no contradictions on that trial. Condition was a marginally significantly predictor of answer correctness on Trial 2, χ^2 (3) = 5.44, p = .072, which was somewhat surprising because there were also no contradictions on this trial. Further examination revealed that the effect was not very pronounced. Compared to the *true-true* condition, learners were significantly less likely to provide a correct answer on Trial 2 in the *false-false* condition, but not in the *true-false* and *false-true* conditions.

A different pattern in the data was observed for contradictory Trials 3 and 4 because condition predicted answer correctness on these trials (χ^2 (3) = 14.4, p = .001 for Trial 3 and χ^2 (3) = 5.35, p = .074 for Trial 4). Learners were significantly less likely to provide correct answers on Trials 3 and 4 in the *true-false* and *false-false* conditions compared to the *true-true* condition. Learners in the *false-true* condition were equally likely to respond accurately as the *true-true* condition for Trial 3 but were less likely to respond accurately for Trial 4. This overall pattern in the data suggests that confusion was highest for the *true-false and false-false* conditions, moderate for the *false-true* condition, and low for the *true-true* condition. Therefore, the delayed contradictions appear to be an effective method to induce uncertainty and presumably confusion in learners.

The results also indicated that learner confusion persisted after reading the explanatory text, because learners in all three experimental conditions were significantly less likely to answer correctly for Trial 5, χ^2 (3) = 32.5, p < .001. Another possibility is that, the confusion was alleviated after reading the texts but was reignited after the Trial 5 contradictions.

4.2.3. Performance on the multiple-choice posttest. Proportional scores on the multiple-choice posttest were regressed on condition with a mixed-effects linear regression model (see Table 6). The overall model was not significant (p = .316). Next, we tested whether self-reported confusion prior to reading the explanatory text moderated the effect of condition on learning by adding the *pre-reading confusion* (not confused vs. confused) × *condition* interaction term to the model. The interaction term was nearly significant, F(3, 296) = 2.09, MSe = .079, p = .051). The interaction was probed by regressing proportional scores on condition (low vs. high) for the confused and not confused cases separately. There were no condition effects on proportional scores when learners were not confused (p = .108). A different pattern was discovered for the cases where learners reported some confusion. Posttest scores for these cases were significantly higher (p = .011) in the *true-false* condition compared to the *true-true* control.

4.2.4. Near and far transfer tests. Each near and far transfer case study was coded as a 1 if a learner correctly detected the flaw or 0 if the flaw was missed. We investigated if condition had an effect on performance on near and far transfer *flaw detection scores* with two mixed-effects logistic regression models. Neither model was significant (p = .134 and .302 for near and far transfer tests, respectively), presumably because this analysis did not segregate cases when a learner was confused from when they were not confused. This was addressed by adding the *pre-reading confusion* × *condition* interaction, but this also did not yield significant models (p > .05).

Since this subjective measure of confusion did not result in any interesting effects, we considered whether the data could be better explained by the inferred measure of confusion. More specifically, we tested whether learners' responses to the forced-choice question in Trial 4 moderated the effect of condition on learning. We focused on Trial 4 because the trialogues associated with this trial are very specific in terms of the flaw with the sample case study and learners are presented with the explanatory text immediately after this trial. The analysis consisted of predicting near and far transfer flaw detection scores from the *Trial 4 answer correctness* (0 = incorrect vs. 1 = correct) × *condition* interaction.

A significant interaction term was discovered for the near transfer model (χ^2 (7) = 17.0, p = .009), while the interaction term was marginally significant for the far transfer model (χ^2 (7) = 11.0, p = .069). We probed these interactions by constructing separate models for cases where learners answered correctly on Trial 4 versus when they answered incorrectly. There was no significant effect of condition on performance on either transfer tests when learners provided correct responses on Trial 4. On the other hand, the results were much more interesting when learners responded incorrectly (see Table 7). Specifically, learners who answered incorrectly were significantly more likely to correctly detect flaws on the near transfer case studies in all three experimental conditions compared to the *true-true* control. These learners were also more likely to detect flaws in the more difficult far transfer problems in the *true-false* and *false-true* condition compared to the *true-true* control. The magnitude of this effect for the far transfer test is impressively large (see Figure 3). For example, learners are 3.7 (B = 1.3; $e^{1.3}$ = 3.7) times more likely to accurately detect a flaw in the *false-true* condition compared to the *true-true* condition when they responded incorrectly on Trial 4.

There was the concern that these effects could merely be attributed to guessing. This could occur if the confused learners simply selected more options as potential flaws. This was addressed by computing the proportion of false-alarms for near and far transfer case studies and regressing false alarms on the *confusion* × *condition* interaction. Fortunately, neither model was significant (p > .05).

4.2.5. Effect of prior knowledge. On average, participants answered 57% (*SD* = 23.4%) of the pretest questions correctly. We tested whether prior knowledge moderated the effect of condition on pre-reading confusion, answer correctness on the five forced-choice questions, on posttest scores, and on performance on the near and far transfer flaw detection tasks. The analyses proceeded by assessing whether the *prior knowledge* × *condition* interaction term significantly predicted these nine dependent variables. The interaction term was not significant for self-reported confusion, answer correctness on Trials 1, 2, and 5, posttest scores, or far transfer flaw detection rates (p > .05). However, the interaction was significant for answer correctness on Trial 3 (χ^2 (7) = 25.3, p < .001), on Trial 4 (χ^2 (3) = 27.7, p < .001), and on flaw detection performance on the near transfer test (χ^2 (7) = 23.7, p < .001).

The interactions were probed by dividing the learners into low and high prior knowledge groups on the basis of a median split and testing the effect of condition for each group separately.

Prior knowledge had the most significant effect on the *false-false* condition. When compared to the *true-true* condition, the low prior knowledge learners were significantly less likely (*B* = -1.68, *p* = .001) to answer correctly on Trial 3 in the *false-false* condition. This effect was not observed for the learners with high prior knowledge (*B* = -.530, *p* = .179). This pattern was replicated in Trial 4. The *false-false* coefficient was significant for low prior-knowledge learners (*B* = -.977, *p* = .017) but was non-significant for their high prior-knowledge counterparts (*B* = -.664 *p* = .171). Interestingly, an opposite pattern was discovered for performance on the near transfer test. The low prior knowledge learners were significantly *more likely* to correctly detect flaws on the near transfer case studies in the *false-false* condition (*B* = 1.09, *p* = .011) compared to the *true-true* control. No significant differences emerged for the high prior knowledge learners (*B* = -.222, *p* = .360).

4.3. Discussion

Experiment 2 attempted to correct identifiable problems with Experiment 1 as well as replicate and extend some of the findings from the earlier experiment. The major differences across experiments involved the use of a delayed contradiction manipulation, inclusion of an online confusion measure in lieu of the offline retrospective affect judgment measure, providing learners with explanatory texts to potentially alleviate their confusion, and the inclusion of a near and far transfer flaw detection task. The results from Experiment 2 both complemented and extended findings from Experiment 1 as discussed in the General Discussion below.

5. General Discussion

5.1. Motivation and Overview of Research

The present research aligns with a recent emphasis on understanding how affective and cognitive processes interact and influence learning (Ainley, Corrigan, & Richardson, 2005; Buff, Reusser, Rakoczy, & Pauli, 2011; Frenzel, Pekrun, & Goetz, 2007; Huk & Ludwigs, 2009;

Jarvenoja & Jarvela, 2005; Pekrun, 2006). However, our present focus was on one particular affective state: confusion. We adopted a theoretical model that posits that learning environments that challenge and potentially confuse learners can be attractive alternatives to more traditional learning activities. This model was inspired by a number of theoretical perspectives including impasse-driven theories of learning (VanLehn et al., 2003), models that highlight the role of cognitive disequilibrium during learning and problem solving (Graesser, Lu, et al., 2005; Graesser & Olde, 2003; Piaget, 1952; Siegler & Jenkins, 1989), models of conceptual change (Chi, 2008; Dole & Sinatra, 1998; Nersessian, 2008), theories that emphasize the importance of interruptions as antecedents of disequilibrium (Mandler, 1976, 1990), and by major statements on the merits of challenges and desirable difficulties during learning (Bjork & Linn, 2006; Linn, Chang, Chiu, Zhang, & McElhaney, in press).

The experiments tested two hypotheses that were derived from the model. The *contradictory information* hypothesis (Hypothesis 1) posits that confusion is triggered when the learner is forced to make a decision but there are contradictions and other anomalies in the available information. However, rather than being detrimental to learning outcomes, the *facilitative confusion* hypothesis (Hypothesis 2) states that confusion can help learning because it promotes deep inquiry and effortful deliberation, on the condition that learners have the ability to appropriately manage their confusion or additional pedagogical scaffolds are available. This is because confusion that is caused by interruptions and other deviations from expectations leads to a restructuring of the cognitive system (D'Mello, Dale, & Graesser, 2012; Piaget, 1952), thereby making the system more attuned to receive information, critically evaluate it, and process it more deeply.

We tested these hypotheses in two experiments that used (a) contradictory information (Experiment 1) and delayed contradiction (Experiment 2) manipulations to induce confusion, (b) offline retrospective affect judgment (Experiment 1 only) and online self-report (Experiment 2 only) protocols as self-report measures of confusion, (c) performance on embedded post-contradiction questions as an objective indicator of confusion, (d) explanations at critical points to correct erroneous information and help externally-regulate confusion (both experiments), (e) an explanatory text intervention to help learners self-regulate confusion (Experiment 2 only), (f) multiple-choice knowledge tests to measure prior knowledge and learning (Experiments 1 and 2), and (g) a near and far transfer flaw detection task to assess knowledge transfer (Experiment 2 only).

The results were illuminating in a number of respects. The major findings were that (a) the contradictions were effective in inducing confusion, (b) self-reports of confusion were not very sensitive to the manipulations, (c) a more objective and indirect measure consisting of learners' responses to forced-choice questions following contradictions was more sensitive at inferring confusion, (d) there were no identifiable learning effects when learners were not confused, (e) learners who were confused by the contradictions were much more likely to perform better on multiple-choice posttests and on tests of knowledge transfer, and (f) prior knowledge had small moderation effects on confusion and learning. The subsequent discussion focuses on aligning these major findings with the two hypotheses, discusses limitations and future directions, and considers the theoretical and applied implications of our work.

5.2. Alignment of Findings with Hypotheses

The results from both experiments supported the *contradictory information* hypothesis (Hypothesis 1) in that the contradictions were successful in inducing states of uncertainty and confusion. One caveat was that the measure of confusion governed the extent to which this hypothesis was supported. If self-reports are considered to be the gold standard to measure emotions then one is forced to conclude that the contradictory information hypothesis was only partially supported (*true-false* condition) in Experiment 1 and not supported in Experiment 2. On the other hand, the *contradictory information hypothesis* (Hypothesis 1) is strongly supported if lower performance on the probing questions that immediately follow contradictions is taken to be an objective indicator of confusion. The idea here is that learners who are uncertain about how to proceed will sometimes side with the correct opinion and other times side with the incorrect opinion. This increased variance in their responses is presumably caused by their underlying confusion and will be reflected in an overall lower accuracy score on these probing questions. The fact that learner responses to these questions systematically degraded in a manner that was dependent upon the source and severity of the contradictions provides some support for this view. Specifically, learner responses were highly accurate when both agents were correct and there were no contradictions. But response quality decreased and uncertainty increased when one agent was incorrect (*true-false* and *false-true*). Uncertainty was theoretically at maximum when both agents were incorrect, even though they did not contradict each other (*false-false*). But this depended on the learners' prior knowledge as discussed in more detail below.

The discovery that levels of confusion moderated the effect of condition on learning in both experiments provided considerable empirical support for the *facilitative confusion* hypothesis (Hypothesis 2). The results indicated that learners who were confused by the manipulation in the *true-false* (Experiments 1 and 2) and *false-false* (Experiment 1 only) conditions demonstrated significantly higher learning gains than those in the no-contradiction control. Importantly, this effect was also observed for both the near and far transfer tests in Experiment 2. Learners who were confused by the contradictions were significantly more accurate at detecting flaws in transfer case studies in the experimental conditions (all three conditions for near transfer; only *true-false* and *false-true* for far transfer) than the control condition that did not include contradictions. Taken together, our results support the claim that confusion can be effective in promoting deeper processing and thereby increased learning gains.

We also expected prior-knowledge to moderate the impact of the contradictions on confusion and learning, but this was largely unconfirmed. One exception was that the low prior knowledge learners were less likely to provide correct responses to the forced-choice questions in the *false-false* condition. The effect was also only found in Experiment 2, where an initial sense of agreement on one opinion was violated by agreement on an opposing opinion. This prior-knowledge effect in the *false-false* condition might be explained by the fact that the discrepancy was more implicit in this condition since there was incorrect information but without overt contradictions. The more knowledgeable learners might have sensed that something was awry when both agents switched from a correct to an incorrect opinion. On the other hand, the low prior-knowledge learners might have simply sided with the incorrect opinion that was asserted by both agents. They might have later realized that their opinion was incorrect when the tutor agent asserted the correct information at the end of the trialogue. This would explain why these learners had better performance (compared to the control condition) on the near transfer test in Experiment 2.

5.3. Limitations and Future Directions

There are three important limitations with this research that need to be addressed in subsequent research activities. First, critics might object to the manipulations on the grounds that intentionally confusing learners by providing misleading information and contradictions is not in the best interests of the learner. We acknowledge this and similar reactions to the manipulations, but should point out that: (a) any misleading information transmitted in a learning session was corrected at the end of that session, (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 in the present study were consenting research participants, and (e) participants were fully debriefed at the end of the experiment. In reference to the second point about learning gains, the data actually showed an opposite pattern in that learners who were confused by the contradictions learned more than those who were not confused. It should also be noted that some instructors design problems and test items that attempt to confuse learners by intentionally leaving out information, providing conflicting alternatives on multiple-choice tests, and increasing task difficulty. These strategies are somewhat similar to the present manipulations with the exception that we directly acknowledge that the end goal is to productively confuse students.

The second limitation with the study pertains to the lack of sensitivity of the self-report measures. Self-reports are an attractive measure of emotion because they are relatively easy to administer and can be interpreted at face value. However, their validity depends upon a number of factors that are outside of the control of the researcher. Some of these include the learners' honesty and willingness to report their emotions (Rasinski, Visser, Zagatsky, & Rickett, 2005), their resilience to social pressures when it comes to reporting negative emotions such as confusion and frustration (Tourangeau & Yan, 2007), the accuracy of learners' introspection in terms of possessing the requisite emotional intelligence to correctly label their emotions (Goleman, 1995), and the requirement that the emotional episode is sufficiently pronounced to make it to learners' consciousness so that it can be subjectively accessed (Rosenberg, 1998). To illustrate some of these validity concerns of self-reports, Figure 4 shows several examples of confused faces obtained over the course of the learning sessions in both experiments. Some of the facial displays were not accompanied by self-reports of confusion, although the known indicators of confusion (lowered brow and tightened lids, or Action Units 4 and 7, respectively)

(Craig, D'Mello, Witherspoon, & Graesser, 2008; Grafsgaard, Boyer, & Lester, 2011; McDaniel et al., 2007) are clearly visible on the face.

It is difficult to identify whether any of the specific factors listed above contributed to the lack of sensitivity of the self-reports in the present study. What is clear, however, is that future research should consider alternate methods to investigate emotions in addition to self-reports. Some possibilities include online observations by external judges (Rodrigo & Baker, 2011), retrospective coding of videos by trained judges (D'Mello & Graesser, 2012), or even physiological and behavioral instrumentation (Arroyo et al., 2009; Calvo & D’Mello, 2010). Perhaps the most defensible position is to include two or more additional measures in addition to the learner self-reports so that the data can be triangulated.

The third limitation is related to the generally weak effects of prior knowledge on the dependent variables. This can be attributed to two factors. A vast majority (> 80%) of the participants had not taken a course in research methods and statistics, so the sample was rather homogenous in terms of previous exposure to scientific reasoning. Furthermore, the pretest, which consisted of eight relatively simple definition questions, might not have been sufficiently diagnostic of learner prior-knowledge. A relatively short and shallow pretest was selected because there was the concern that a more comprehensive pretest might cue learners in to the source of contradictions, which would potentially mitigate the impact of the contradictory information manipulation. Nevertheless, there is the need to test for prior-knowledge effects more systematically. This can be accomplished by contrasting novice learners who have not completed a research methods or statistics course with more advanced learners and by including a more diagnostic test of prior knowledge. This is an important direction for future work.

5.4. Theoretical Implications

The importance of disequilibrium, impasses, dissonance, and conflict in learning and problem solving has a long history in psychology that spans the developmental, educational, social, and cognitive sciences (Berlyne, 1978; Chinn & Brewer, 1993; Collins, 1974; Festinger, 1957; Graesser & Olde, 2003; Laird, Newell, & Rosenbloom, 1987; Limón, 2001; Miyake & Norman, 1979; Mugny & Doise, 1978; Piaget, 1952; Schank, 1999). The notion that these states extend beyond cognition and into emotions has also been acknowledged and investigated for decades (Festinger, 1957; Graesser, Lu, et al., 2005; Lazarus, 1991; Mandler, 1976; Piaget, 1952). The present research advances these theories by highlighting the critical role of confusion in driving deep learning and inquiry.

Our theoretical approach and findings are also consistent with several aspects of the cognitive-affective theory of learning with multimedia (CATLM) (Moreno, 2005; Moreno & Mayer, 2007). CATLM builds upon and extends Mayer's cognitive theory of multimedia learning (Mayer, 2003, 2005) to include diverse media such as virtual reality and agent-based learning environments. The theory is too broad to be comprehensively described in this article, but we have identified some ways that there is an alignment between CATLM and our key assumptions and findings. First, CATLM distinguishes between the information acquisition and the knowledge construction views of learning. The former focuses on the mere addition of information to memory while the latter is concerned with the active integration of new information into knowledge structures (or mental models). Similar to CATLM, the present research is consistent with the knowledge construction view of learning. Second, CATLM claims that the three key processes that underlie deep learning include the selection of relevant information to attend to, mentally organizing the attended information into coherent units, and integrating the newly organized knowledge chunks into existing knowledge structures. This is precisely the perspective of complex learning that we have adopted. A major goal of the present research was to explore how confusion influences these cognitive processes. In particular, confusion focuses attention on discrepant events, it signals a need to initiate effortful deliberation and problem solving processes, and it influences knowledge restructuring when impasse resolution or misconception correction lead to the reorganization of an incomplete or faulty mental model. Third, CATLM posits that metacognitive factors influence learning by mediating cognitive and affective processes. Along these lines, and consistent with Mandler's interruption (discrepancy) theory (Mandler, 1990), the affective state of confusion signals that a discrepancy has been detected in the course of information processing. This signal *interrupts* the processing stream and can make the learner metacognitively aware of the state of his or her knowledge. This can lead to more top-down processing via a conscious recruitment of resources.

The present research also contributes to the refinement and expansion of some existing theories that link affect and cognition. As expressed earlier, states of cognitive disequilibrium and cognitive dissonance have been investigated for several decades. Confusion is usually implicitly implicated by these theories, yet most theoretical frameworks do not directly address this emotion. Most theories also fail to address the temporal dynamics of confusion, even though this is arguably the most interesting aspect to consider because of the highly fluid and ephemeral nature of confusion. An important next step is to extend these theories by considering the chronometry of confusion and its related processes. We have therefore sketched a model that predicts specific confusion trajectories based on the severity of the discrepancy and the results of effortful confusion regulation processes.

The model assumes that individuals encounter discrepancies at multiple levels as they attempt to assimilate incoming information into existing mental models. There is some threshold T_a that needs to be exceeded before the individual is confused. Discrepancies that are not severe enough to exceed T_a are not detected by the individual and there is no confusion. Sometimes the severity of the discrepancy greatly exceeds T_a and the individual is bewildered or flustered. This threshold can be denoted as $T_{b.}$ A moderate level of confusion is experienced when the severity of the discrepancy meets or exceeds T_a but is less than T_b. However, the individual may not elect to attend to the confusion and shifts attentional resources elsewhere. When this occurs, confusion is alleviated very quickly and the length of confusion is less than duration D_a. If the length of the confusion episode exceeds D_a, then the individual has begun to attempt to identify the source of the discrepancy in order to resolve the confusion. When confusion resolution fails and the individual is confused for a long enough duration D_b, then there is the risk of frustration. With a longer duration D_c, there is a persistent frustration, and the risk of disengagement and boredom (i.e., the learner gives up). There is potentially a *zone of optimal confusion* which occurs when: *discrepancy* > T_a and *discrepancy* < T_b and *duration* > D_a and *duration* < D_b.

Recent research that has identified bi-directional *confusion-engagement*, *confusion-frustration,* and *frustration-boredom* transitions provides some evidence in support of this model (see Figure 5) (D'Mello & Graesser, 2012). The confusion-engagement transition is presumably linked to experiencing discrepancies (engagement to confusion) and successfully resolving the confusion (confusion to engagement). The confusion-frustration transition likely occurs when a learner experiences failure when attempting to resolve an impasse (confusion to frustration) and experiences additional impasse(s) when frustrated (frustration to confusion). Transitions involving boredom and frustration are presumably related to a state of being stuck due to persistent failure to the point of disengaging (frustration to boredom) and annoyance from being forced to persist in the task despite having mentally disengaged (boredom to frustration; for possible transitions to anxiety and hopelessness, see Pekrun, 2006).

What is currently missing is systematically specifying and testing the various durations and thresholds of the model, which obviously depend on some interaction between the individual characteristics of the learner and the complexity of the task. Systematically fitting these parameters in a manner that is sensitive to constraints of the learner, the environment, and their interaction is an important item for future work. In addition, it is an important step towards the long-term goal of identifying *zones of optimal confusion* for individual learners.

5.5. Applied Implications

The present results are significant because they constitute some of the first experimental evidence on the advantage of inducing confusion during learning. The most obvious implication of this research is that there might be some practical benefits for designing educational interventions that intentionally perplex learners. Learners complacently experience a state of low arousal when they are in comfortable learning environments involving passive reading and accumulating shallow facts without challenges. However, these comfortable learning environments rarely lead to deep learning. In contrast, deep learning is expected to be higher in environments that present challenges to inspire deep inquiry, provided that the learners attend to impasses and the resultant confusion. Learners also need to have the requisite knowledge and skills to resolve the confusion or alternatively the learning environment needs to provide appropriate scaffolds to help with the confusion resolution process.

The discussion above lends itself to the question of how confusion induction interventions can be deployed in real world educational settings. As a precursor to presenting some of our ideas along this front, we acknowledge that the notion of promoting learning (specifically conceptual change) by interventions that induce cognitive conflict has a rich history in educational psychology (e.g., Chinn & Brewer, 1993; Dreyfus, Jungwirth, & Eliovitch, 1990;

Mason, 2001). Unfortunately, the results of testing these interventions in the classroom have not been very promising, as pointed out in a relatively recent review of this literature (Limón, 2001).

Limón (2001) suggests that one explanation for the lack of impressive effects is that the interventions often fail to promote *meaningful* conflict in the minds of the learners. This is because most of the interventions have primarily focused on learners' cognitive processes, while ignoring individual differences in motivation orientations, prior knowledge, learning styles, values and attitudes about learning, epistemological beliefs, and reasoning abilities. The role of social interactions and peer collaboration has also received less attention, which is unfortunate because classroom learning is inherently a social phenomenon. Simply put, the one-size-fits-all strategy that one is forced to adopt in the classroom makes it extremely difficult to develop an intervention that is likely to induce meaningful conflict in a majority of the learners. What is needed, are interventions that promote conflict in a manner that is aligned with the needs, goals, and abilities of individual learners.

In our view, advanced learning technologies that deliver individualized one-on-one interaction can alleviate several of the concerns highlighted by Limón (2001). Indeed, the multimedia collaborative learning environment in the present study implements some of the features deemed important to induce meaningful conflict in learners. First, the case studies that we selected focused on topics that were expected to be somewhat interesting for our sample of college students (e.g., discussions on the importance of buying textbooks, the efficacy of diet pills, placebo effects within the context of alcohol consumption, etc.). Second, there were multiple rounds of contradictory-information trials, thereby providing multiple opportunities for the interventions to have an effect. Third, the contradictions were embedded within the primary learning activity, so there was some immediate relevance. In fact, attending to and processing the contradictions was the primary activity. Fourth, learners had to provide a response immediately following each contradiction, thereby further increasing the likelihood that they would attend to the contradictory information. Fifth, scaffolds to help learners process the contradictions were provided throughout the system via the case studies being displayed on the screen, the scrolling dialogue history, the explanatory text provided in Experiment 2, and the explanations provided by the tutor at key junctions in the conversation. Sixth, the entire learning activity was embedded within a collaborative learning context involving agents with well-defined roles, thereby simulating some of the social aspects of learning.

In summary, we believe that advanced learning technologies hold considerable promise in increasing learning via cognitive conflict because of their ability to dynamically tailor instruction to individual learners. However, there is much more work that needs to be done before these technologies can be deployed in classroom contexts. There is the need for more basic research on confusion, its antecedents, and consequents. There is the challenge of incorporating automated technologies to sense the induced confusion so that the system can incorporate this information while planning its next move (D'Mello & Graesser, 2010). There might also be some benefit to encouraging learners to provide self-explanations at critical points during the trialogues. This creates the need for automated systems to evaluate these natural language responses (Lehman, Mills, D'Mello, & Graesser, in press). Finally, it is important that appropriate scaffolds are implemented in order to help learners intelligently manage their confusion.

The question is sometimes raised as to whether there are ethical issues that arise from interventions that promote cognitive conflict by planting false information. One solution to this problem is to utilize confusion induction methods that do not transmit any false information to the learners. This is a viable solution because there are several antecedents of confusion which can be used in lieu of a false-information manipulation (D'Mello & Graesser, in press; Silvia, 2010). For example, earlier we described a method that successfully creates confusion with breakdown scenarios (see Introduction). This method was quite effective in inducing confusion and is likely to pass ethical muster because it does not involve any false information.

There is also the manner of identifying *who* might benefit from a confusion induction intervention. It is probably not a very sensible strategy to attempt to confuse a struggling learner or to induce confusion during high stakes learning activities, at least until confusion induction techniques are refined and their consequences are better understood. Currently, these interventions are ideally suited for gifted learners who are often bored and disengage from learning activities due to a lack of challenge (Pekrun, Goetz, Daniels, Stupnisky, & Perry, 2010). There is also a risk of confusing students who are cautious learners instead of academic risk takers (Clifford, 1988; Meyer & Turner, 2006) or learners who have a fixed (entity) instead of growth (incremental) mindset of intelligence (Dweck, 2006). Confusion interventions are best for adventurous learners who want to be challenged with difficult tasks, are willing to risk failure, and manage negative emotions when they occur because they understand that failure is an inevitable part of a successful path towards proficiency development. These learners can be challenged at the extremes of their zones of proximal development (Brown, Ellery, & Campione, 1998; Vygotsky, 1986), provided that appropriate scaffolds are in place when they struggle or they can manage the challenges with self-regulated learning strategies.

The interventions might also be suitable for passive learners with moderate skills, motivations, and academic ambitions. Interventions that confuse these complacent learners via contradictions, incongruities, anomalies, system breakdowns, and difficult decisions might be just what is needed to jolt them out of their perennial state of passively receiving information and inspire them to focus attention, engage fully, think more deeply, and learn for mastery.

Appendix

Sample questions on knowledge tests (correct answers are bolded)

1. (*Definition question*) Random assignment refers to:
    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.

2. (*Function question*) Random assignment is important because:
    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.
    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.

3. (*Example question*) Which of the following studies is the best example of random assignment of participants to groups?
    a. a researcher wants to study the impact of class size on test performance so he chooses a 300-student introduction to psychology class from one university and a 30-student class from another university to participate in the study.
    b. a researcher wants to assess the impact of time of day on learning so she uses a coin flip to place students in either the day or evening experimental session.
    c. a researcher wants to assess the impact of a fertilizer on plant growth so she provides farmer Brown’s field with the fertilizer and nothing for farmer Jones’s field.
    d. a researcher wants to assess the impact of exposure to vitamin B12 on the immune system so he recruits patients from one clinic that recommends B12 and patients from another clinic that does not recommend it.

References

Ainley, M., Corrigan, M., & Richardson, N. (2005). Students, tasks and emotions: Identifying the contribution of emotions to students' reading of popular culture and popular science texts. *Learning and Instruction, 15*(5), 433-447. doi: 10.1016/j.learninstruc.2005.07.011

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. Graesser (Eds.), *Proceedings of 14th international conference on artificial intelligence in education* (pp. 17-24). Amsterdam: IOS Press.

Bates, D. M., & Maechler, M. (2010). Lme4: Linear mixed-effects models using s4 classes. Retrieved from http://CRAN.R-project.org/package=lme4

Berlyne, D. (1978). Curiosity in learning. *Motivation and Emotion, 2*, 97-175. doi: 10.1007/BF00993037

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56-64). New York: Worth Publishers.

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, A., Ellery, S., & Campione, J. (1998). Creating zones of proximal development electronically. In J. Greeno & S. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 341-367). Mahwah, NJ: Lawrence Erlbaum.

Brown, J., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science, 4*(379-426,). doi: 10.1016/S0364-0213(80)80010-3

Buff, A., Reusser, K., Rakoczy, K., & Pauli, C. (2011). Activating positive affective experiences in the classroom: "Nice to have" or something more? *Learning and Instruction, 21*(3), 452-466.

Calvo, R. A., & D'Mello, S. K. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18-37. doi: 10.1109/T-AFFC.2010.1

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. doi: 10.1016/S0020-7373(88)80050-6

Chi, M. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 61-82). New York: Routledge.

Chinn, C., & Brewer, W. (1993). The role of anomalous data in knowledge acquisition - a theoretical framework and implications for science instruction. *Review of Educational Research, 63*(1), 1-49. doi: 10.2307/1170558

Clifford, M. (1988). Failure tolerance and academic risk-taking in ten- to twelve-year-old students. *British Journal of Educational Psychology, 58*(15-27). doi: 10.1111/j.2044-8279.1988.tb00875.x

Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. *Trends in Cognitive Sciences, 11*(9), 393-399. doi: 10.1016/j.tics.2007.08.005

Collins, A. (1974). Reasoning from incomplete knowledge. *Bulletin of the Psychonomic Society, 4*, 254-254.

Craig, S., D'Mello, S., Witherspoon, A., & Graesser, A. (2008). Emote aloud during learning with autotutor: Applying the facial action coding system to cognitive-affective states during learning. *Cognition & Emotion, 22*(5), 777-788.

Craig, S., Graesser, A., Sullins, J., & Gholson, J. (2004). Affect and learning: An exploratory look into the role of affect in learning. *Journal of Educational Media, 29*, 241-250. doi: 10.1080/1358165042000283101

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684. doi: 10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1972). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268-294. doi: 10.1037//0096-3445.104.3.268

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

D'Mello, S., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-adapted Interaction, 20*(2), 147-187.

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*, 145-157. doi: 10.1016/j.learninstruc.2011.10.001

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

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

D’Mello, S., Dale, R., & Graesser, A. (2012). Disequilibrium in the mind, disharmony in the body. *Cognition & Emotion, 26*(2), 362-374. doi: 10.1080/02699931.2011.613668

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. (in press). Emotions during learning with autotutor. In P. Durlach & A. Lesgold (Eds.), *Adaptive technologies for training and education*. Cambridge, U.K: Cambridge University Press.

Darwin, C. (1872). *The expression of the emotions in man and animals*. London: John Murray.

Dimant, R. J., & Bearison, D. J. (1991). Development of formal reasoning during successive peer interactions. *Developmental Psychology, 27*(2), 277. doi: 10.1037//0012-1649.27.2.277

Dole, J. A., & Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist, 33*(2/3), 109-128. doi: 10.1207/s15326985ep3302&3_5

Dreyfus, A., Jungwirth, E., & Eliovitch, R. (1990). Applying the “cognitive conflict” strategy for conceptual change—some implications, difficulties, and problems. *Science Education, 74*(5), 555-569. doi: 10.1002/sce.3730740506

Dweck, C. (2006). *Mindset*. New York: Random House.

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

Forbes-Riley, K., & Litman, D. (2009). Adapting to student uncertainty improves tutoring dialogues. In V. Dimitrova, R. Mizoguchi & B. Du Boulay (Eds.), *Proceedings of the 14th international conference on artificial intelligence in education* (pp. 33-40). Amsterdam: IOS Press.

Forbes-Riley, K., & Litman, D. (2010). Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech and Language, 25*(1), 105-126. doi: http://dx.doi.org/10.1016/j.csl.2009.12.002

Forbes-Riley, K., & Litman, D. J. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication, 53*(9-10), 1115-1136. doi: 10.1016/j.specom.2011.02.006

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Perceived learning environment and students' emotional experiences: A multilevel analysis of mathematics classrooms. [Article]. *Learning and Instruction, 17*(5), 478-493. doi: 10.1016/j.learninstruc.2007.09.001

Goleman, D. (1995). *Emotional intelligence*. New York: Bantam Books.

Graesser, A., Chipman, P., Haynes, B., & Olney, A. (2005). Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48*(4), 612-618. doi: 10.1109/TE.2005.856149

Graesser, A., 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* (pp. 569-571). Amsterdam: IOS Press.

Graesser, A., Lu, S., Olde, B., 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. doi: 10.3758/BF03193225

Graesser, A., & Olde, B. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology, 95*(3), 524-536. doi: 10.1037/0022-0663.95.3.524

Graesser, A., Ozuru, Y., & Sullins, J. (2010). What is a good question? In M. McKeown & G. Kucan (Eds.), *Bringing reading research to life* (pp. 112-141). New York: Guilford.

Grafsgaard, J., Boyer, K., & Lester, J. (2011). Predicting facial indicators of confusion with hidden markov models. In S. D'Mello, A. Graesser, B. Schuller & J. Martin (Eds.), *Proceedings of the 4th international conference on affective computing and intelligent interaction (acii 2011)* (pp. 97-106). Berlin Heidelberg: Springer.

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

Huk, T., & Ludwigs, S. (2009). Combining cognitive and affective support in order to promote learning. *Learning and Instruction, 19*(6), 495-505.

Izard, C. (2010). The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review, 2*(4), 363-370. doi: 10.1177/1754073910374661

Jarvenoja, H., & Jarvela, S. (2005). How students describe the sources of their emotional and motivational experiences during the learning process: A qualitative approach. *Learning and Instruction, 15*(5), 465-480.

Keltner, D., & Shiota, M. (2003). New displays and new emotions: A commentary on rozin and cohen (2003). *Emotion, 3*(86-91). doi: 10.1037/1528-3542.3.1.86

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar - an architecture for general intelligence. *Artificial Intelligence, 33*(1), 1-64. doi: 10.1016/0004-3702(87)90050-6

Lazarus, R. (1991). *Emotion and adaptation*. New York: Oxford University Press.

Lehman, B., D'Mello, S., & Person, N. (2010). The intricate dance between cognition and emotion during expert tutoring. In J. Kay & V. Aleven (Eds.), *Proceedings of 10th international conference on intelligent tutoring systems* (pp. 433-442). Berlin/Heidelberg: Springer.

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 the 9th international conference on intelligent tutoring systems* (pp. 50-59). Berlin, Heidelberg: Springer.

Lehman, B., Mills, C., D'Mello, S., & Graesser, A. (in press). Automatic evaluation of learner self-explanations and erroneous responses for dialogue-based itss *Proceedings of the 11th international conference on intelligent tutoring systems*.

Limón, M. (2001). On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and Instruction, 11*(4–5), 357-380. doi: 10.1016/s0959-4752(00)00037-2

Linn, M. C., Chang, H., Chiu, J., Zhang, Z., & McElhaney, K. (in press). Can desirable difficulties overcome deceptive clarity in scientific visualizations? In A. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of robert a. Bjork*.

Mandler, G. (1976). *Mind and emotion*. New York: Wiley.

Mandler, G. (1984). *Mind and body: Psychology of emotion and stress*. New York: W.W. Norton & Company.

Mandler, G. (1990). Interruption (discrepancy) theory: Review and extensions. In S. Fisher & C. L. Cooper (Eds.), *On the move: The psychology of change and transition* (pp. 13-32). Chichester: Wiley.

Mandler, G. (1999). Emotion. In B. M. Bly & D. E. Rumelhart (Eds.), *Cognitive science. Handbook of perception and cognition* (2nd ed., pp. 367-382). San Diego, CA: Academic Press.

Mason, L. (2001). Responses to anomalous data on controversial topics and theory change. *Learning and Instruction, 11*(6), 453-483. doi: 10.1016/s0959-4752(00)00042-6

Mayer, R. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction, 13*(2), 125-139. doi: 10.1016/s0959-4752(02)00016-6

Mayer, R. (Ed.). (2005). *The cambridge handbook of multimedia learning*. New York: Cambridge University Press.

McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., & Graesser, A. (2007). Facial features for affective state detection in learning environments. In D. McNamara & G. Trafton (Eds.), *Proceedings of the 29th annual meeting of the cognitive science society* (pp. 467-472). Austin, TX: Cognitive Science Society.

Meyer, D., & Turner, J. (2006). Re-conceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review, 18*(4), 377-390. doi: 10.1007/s10648-006-9032-1

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

Miyake, N., & Norman, D. (1979). To ask a question, one must know enough to know what is not known. *Journal of Verbal Learning and Verbal Behavior, 18*(3), 357-364. doi: 10.1016/S0022-5371(79)90200-7

Moreno, R. (2005). Instructional technology: Promise and pitfalls. In L. PytlikZillig, M. Bodvarsson & R. Bruning (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 1-19). Greenwich, CT: Information Age Publishing.

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review, 19*(3), 309-326. doi: 10.1007/s10648-007-9047-2

Mugny, G., & Doise, W. (1978). Socio-cognitive conflict and structure of individual and collective performances. *European Journal of Social Psychology, 8*(2), 181-192.

Nersessian, N. (2008). Mental modeling in conceptual change. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 391-416). New York: Routledge.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*(4), 315-341.

Pekrun, R., Goetz, T., Daniels, L., Stupnisky, R. H., & Perry, R. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology, 102*(3), 531-549. doi: 10.1037/a0019243

Pekrun, R., & Stephens, E. J. (2012). Academic emotions. In K. Harris, S. Graham, T. Urdan, S. Graham, J. Royer & M. Zeidner (Eds.), *Apa educational psychology handbook, vol 2: Individual differences and cultural and contextual factors* (pp. 3-31). Washington, DC: American Psychological Association.

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

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

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. doi: 10.1016/j.jesp.2004.07.001

Rodrigo, M., & Baker, R. (2011). 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. 183-202). New York: Springer.

Rosenberg, E. (1998). Levels of analysis and the organization of affect. *Review of General Psychology, 2*(3), 247-270. doi: 10.1037//1089-2680.2.3.247

Roth, K. J., Druker, S. L., Garnier, H. E., Lemmens, M., Chen, C., Kawanaka, T., . . . Gallimore, R. (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.

Rozin, P., & Cohen, A. (2003). High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of maturally occurring facial expressions of americans. *Emotion, 3*, 68-75.

Schank, R. (1999). *Dynamic memory revisited*: Cambridge University Press.

Siegler, R., & Jenkins, E. (Eds.). (1989). *Strategy discovery and strategy generalization*. Hillsdale, NJ,: Lawrence Erlbaum Associates.

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. doi: 10.1037/a0017081

Stein, N., Hernandez, M., & Trabasso, T. (2008). Advances in modeling emotions and thought: The importance of developmental, online, and multilevel analysis. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 574-586). New York: Guilford Press.

Stein, N., & Levine, L. (1991). Making sense out of emotion. In W. Kessen, A. Ortony & F. Kraik (Eds.), *Memories, thoughts, and emotions: Essays in honor of george mandler* (pp. 295-322). Hillsdale, NJ: Erlbaum.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859. doi: 10.1037/0033-2909.133.5.859

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. doi: 10.1207/S1532690XCI2103_01

Vygotsky, L. (1986). *Thought and language*. Cambridge, MA: MIT Press.

Table 1. Sample case study and excerpt of trialogue

Turn	Speaker	Dialogue
There was a study done at a top University where students got the same grade whether they used the textbook or not. In the fall, science students were told that textbooks were optional. For the same class in the spring, students were told that reading the textbook was required. The researchers found no differences on the final exams. So there is no need to buy textbooks. <*Research study*>		
1	Dr. Williams	So we talked while you were reading and Chris thinks that there wasn't anything problematic about this study, but I think there was. <*Tutor & student assertion*>
2	Dr. Williams	Bob, would you not buy textbooks next semester based on this study? Please type buy or not buy. <*Forced-choice question*>
3	Bob	not buy <*Response*>
4	Dr. Williams	We are going to go over our thinking for this study before we come to any final decisions. <*Advance dialogue*>
5	Chris	Well, I think how the participants were put into each condition was good, so that's not a problem. <*Assertion*>
6	Dr. Williams	It was problematic. <*Assertion*>
7	Dr. Williams	Bob, do you think there's a problem with how the participants were put into each group? Please type problem or no problem. <*Forced-choice question*>
8	Bob	problem <*Response*>
9	Dr. Williams	Chris, can the researchers know if the two groups are equivalent? <*Advance dialogue*>
10	Chris	Yes, they can. <*Assertion*>
11	Dr. Williams	I disagree. <*Assertion*>
12	Dr. Williams	Bob, do the researchers know that the two groups are equivalent? Please type know or don't know. <*Forced-choice question*>
13	Bob	don't know <*Response*>
14	Dr. Williams	I think it would've been better if they had randomly assigned people to use or not use the textbook. <*Assertion*>
15	Chris	No, I think signing up for different sections is random enough. <Assertion>
16	Dr. Williams	Bob, should the researchers have used random assignment here? Please type random assignment or no random assignment. <*Forced-choice question*>
17	Bob	no random assignment <*Response*>

Table 2. Proportional occurrence of affective states from Experiment 1

	Proportional Occurrence				**Coefficients (B)**		
Affect	*Tr-Tr*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*
Anxiety	.005	.005	.007	.004	.257	.332	-.297
Boredom	**.349**	**.305**	**.320**	.333	**-.271**	**-.171**	-.082
Confusion	**.076**	**.100**	.098	.081	**.329**	.171	-.073
Curiosity	.085	.081	.093	.085	-.100	.050	-.014
Delight	.018	.013	.015	.018	-.262	-.275	.073
Engaged	**.142**	**.166**	.135	.155	**.261**	-.046	.148
Frustration	.059	.060	.060	.060	.030	.085	.140
Neutral	.260	.262	.261	.256	.005	.042	-.040
Surprise	.006	.008	.011	.008	.336	.646	.193

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

Table 3. Proportion of forced-choice questions correctly answered (Experiment 1)

	Proportion Correct				**Coefficient (B)**		
Trial	*Tr-Tr*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*
Trial 1	.575	.543	.468	.437	-.129	**-.432**	**-.562**
Trial 2	.724	.551	.460	.336	**-.764**	**-1.12**	**-1.64**
Trial 3	.722	.598	.413	.304	**-.573**	**-1.35**	**-1.83**
Trial 4	.696	.595	.440	.352	**-.445**	**-1.07**	**-1.45**
Mean	.679	.572	.445	.357			

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

Table 4. Proportion of correct responses on the posttest (Experiment 1)

	Proportion Correct				**Coefficient (B)**		
Effect	*Tr-Tr*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*
Main Effect							
Condition	.364	.381	.388	.398	.017	.034	.045
Confusion × Condition							
Low	.371	.332	.402	.337	-.028	.051	.005
High	**.356**	**.418**	.359	**.442**	**.090**	.031	**.098**

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

Table 5. Proportion of forced-choice questions correctly answered (Experiment 2)

	Proportion Correct				**Coefficient (B)**		
Trial	*Tr-Tr*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*
No Contradictions							
Trial 1	.750	.789	.724	.789	.228	-.150	.189
Trial 2	**.724**	.711	.737	**.842**	-.108	.054	**.786**
Pre-Reading							
Trial 3	**.487**	**.303**	.474	**.289**	**-1.10**	-.138	**-1.15**
Trial 4	**.697**	**.592**	**.579**	**.539**	**-.596**	**-.673**	**-.805**
Post-Reading							
Trial 5	**.842**	**.592**	**.645**	**.474**	**-1.82**	**-1.51**	**-2.45**

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

Table 6. Performance on the multiple-choice posttest (Experiment 2)

	Proportion Correct				**Coefficient (B)**		
Effect	*Tr-Tr*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*
Main Effect							
Condition	.471	.471	.442	.487	.005	-.025	.016
Confusion × Condition							
Not Confused	.508	.457	.430	.502	-.045	-.071	-.012
Confused	**.341**	**.525**	.487	.435	**.178**	.055	.056

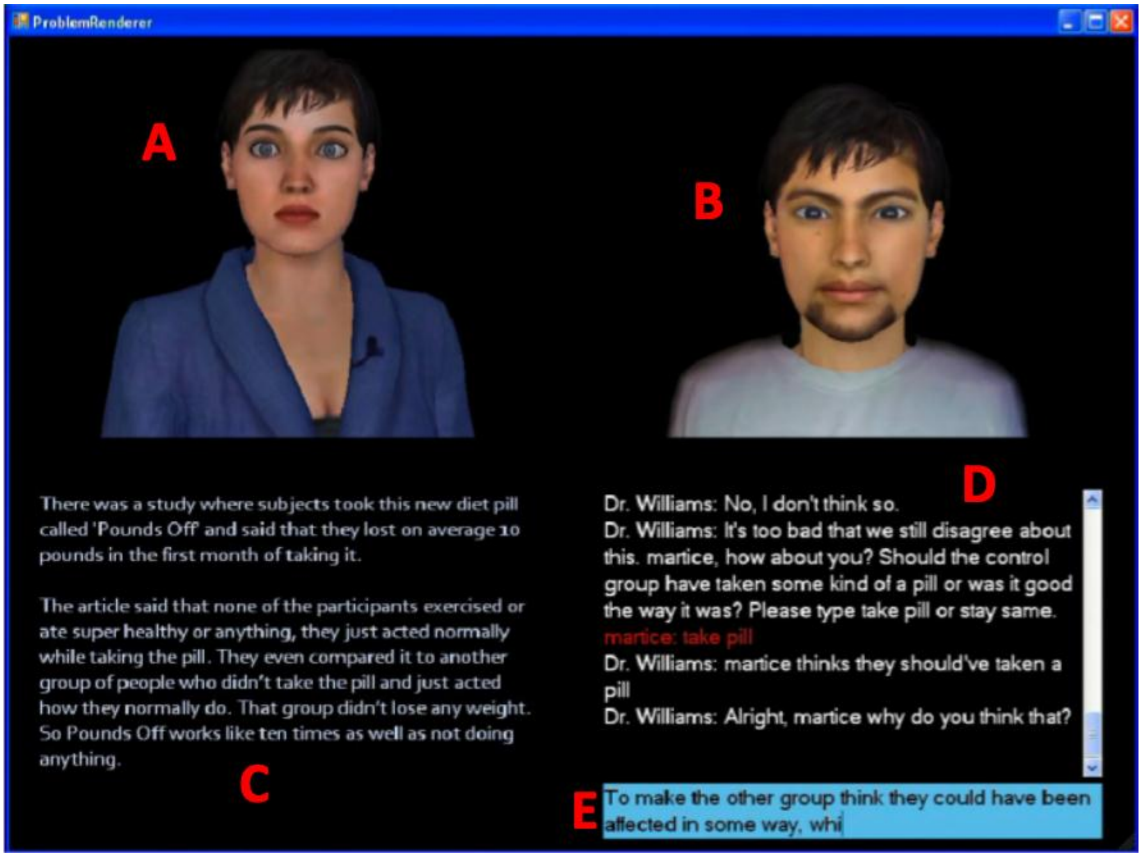
*Note*s. Tr: True; Fl: False; Tr-Tr was the reference group for the models, hence, coefficients for this condition are not shown in the table. Bolded cells refer to significant effects at $p < .05$.

Table 7. Performance on near and far transfer tests (Experiment 2)

	Proportion Correct						**Coefficient (B)**
Transfer	*Tr-Tr*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*	*Tr-Fl*	*Fl-Tr*	*Fl-Fl*
Near Transfer							
Trial 4 Correct	.642	.556	.659	.659	-.339	.019	-.033
Trial 4 Incorrect	**.174**	**.419**	**.469**	**.543**	**1.58**	**1.90**	**2.39**
Far Transfer							
Trial 4 Correct	.264	.267	.159	.220	-.124	-.808	-.443
Trial 4 Incorrect	**.217**	**.387**	**.406**	.257	**1.19**	**1.30**	.275

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

Figure 1. Interface for learning sessions



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

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

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

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

Figure 2. Condition × confusion interaction for performance on the posttest (Experiment 1)

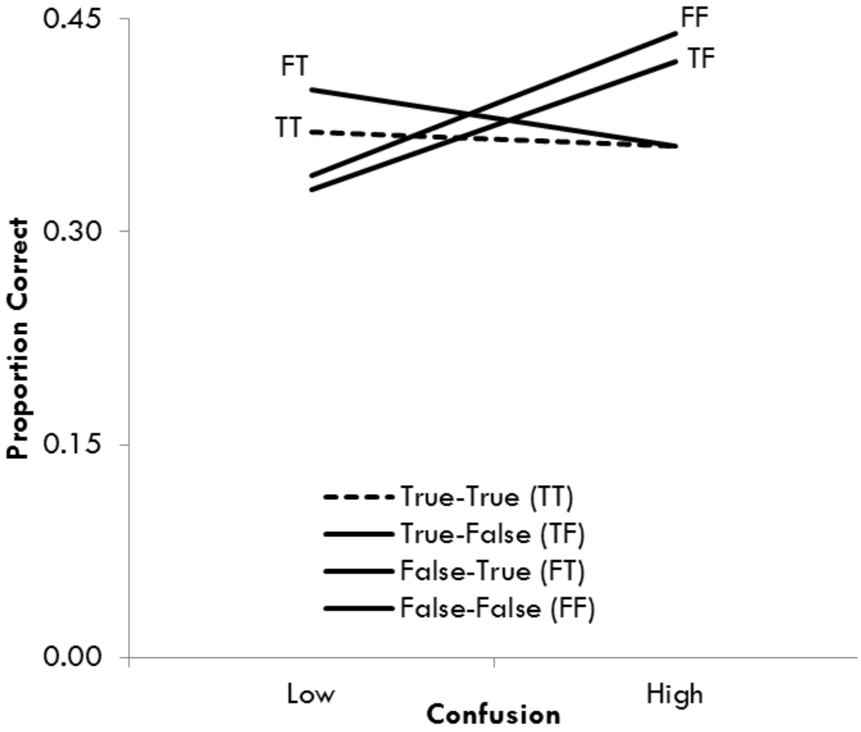


Figure 3. Condition × Trial 4 interaction for performance on far transfer test (Experiment 2)

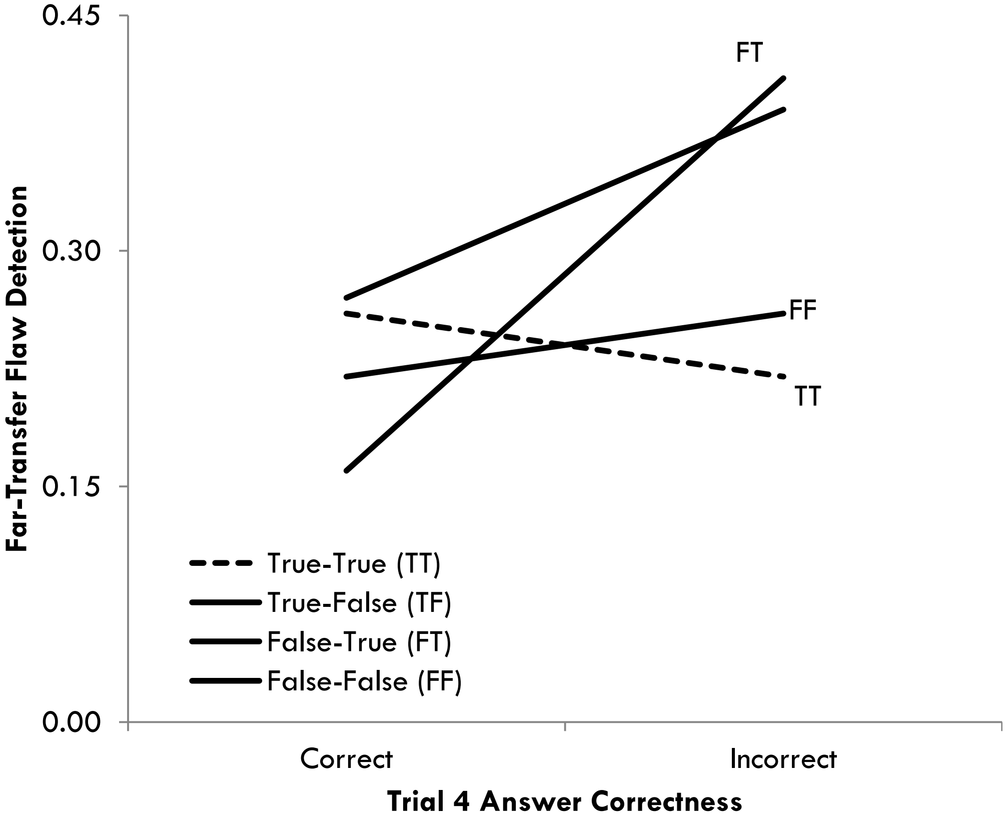


Figure 4. Examples of confused faces from participants in Experiments 1 and 2

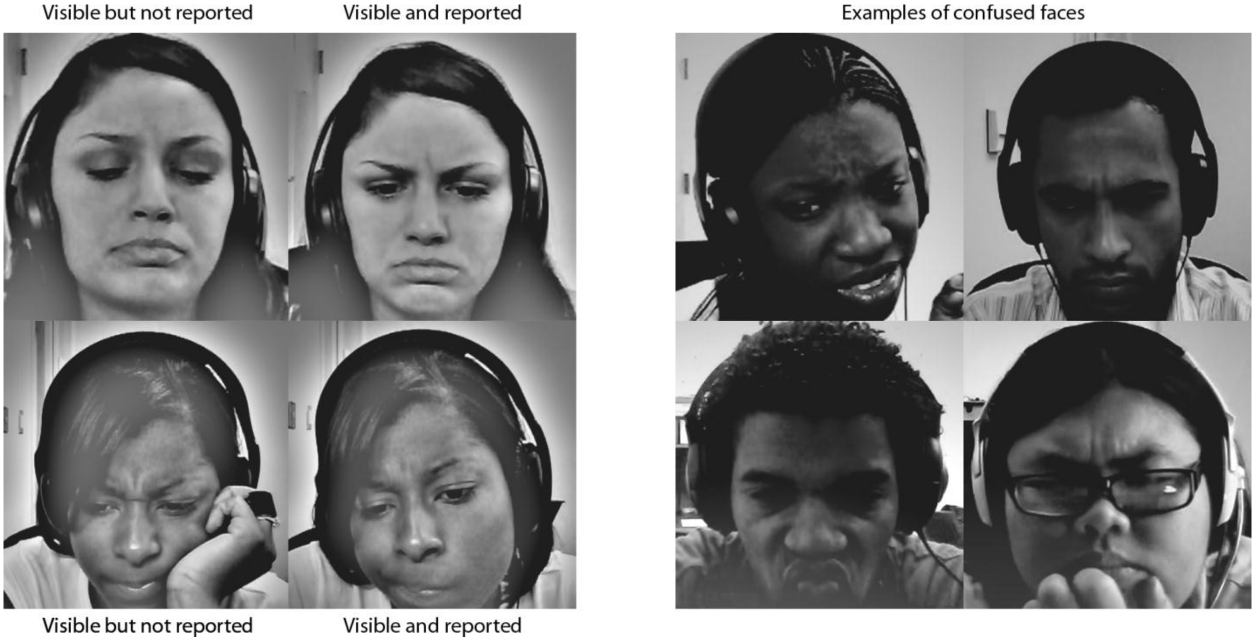
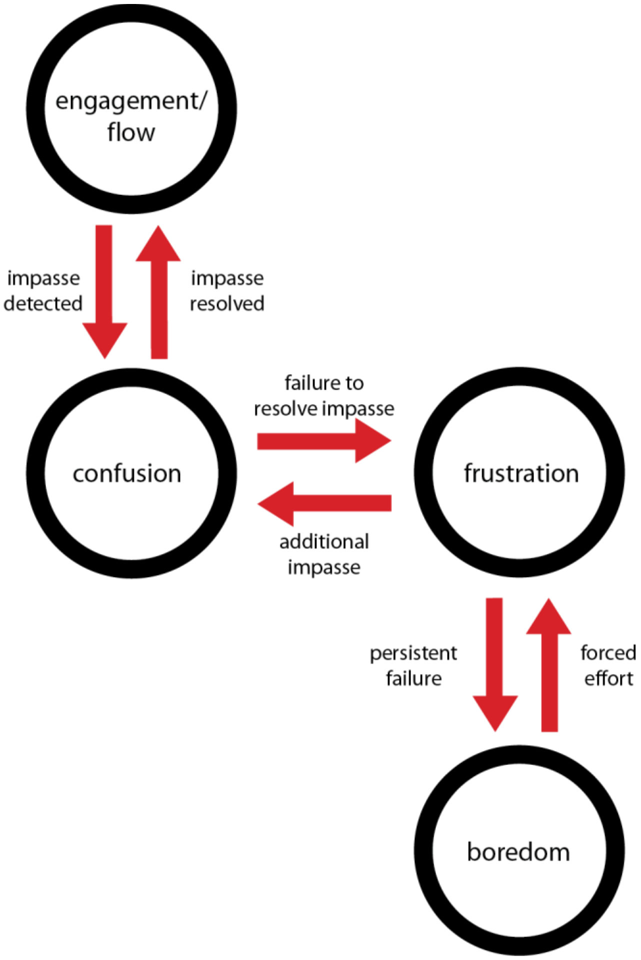


Figure 5. Observed emotion transitions and their hypothesized causes (Image adapted from D’Mello & Graesser (2012))


engagement/ flow
impasse detected
impasse resolved
confusion
failure to resolve impasse
frustration
additional impasse
persistent failure
forced effort
boredom