7-18-2014

Interventions to Regulate Confusion during Learning

Blair Allison Lehman

INTERVENTIONS TO REGULATE CONFUSION DURING LEARNING

by

Blair Allison Lehman

A Dissertation

Submitted in Partial Fulfillment of the

Requirement for the Degree of

Doctor of Philosophy

Major: Psychology

The University of Memphis

August 2014

Acknowledgements

The current dissertation thanks the research colleagues in the Emotive Computing Group and the Tutoring Research Group (TRG) at the University of Memphis and the University of Notre Dame (http://emotion.autotutor.org). Special thanks to Victoria Maher, Fadumo Nur, and Eliana Silbermann for data collection. Special thanks are also needed for Sidney D'Mello, Art Graesser, and Natalie Person for providing invaluable guidance and support from the beginning of this project.

This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1108845), the Institute of Education Sciences (IES, U.S. Department of Education (DoE), through Grant R305A080594, and the U.S. Office of Naval Research (N00014-05-1-0241). Any opinions, findings, and conclusions, or recommendations expressed in this dissertation are those of the author and do not necessarily reflect the views of the NSF, IES, DoE, or ONR.

Abstract

Lehman, Blair Allison. PhD. The University of Memphis. August 2014. Interventions to Regulate Confusion during Learning. Major Professor: Arthur Graesser, PhD.

Confusion provides opportunities to learn at deeper levels. However, learners must put forth the necessary effort to resolve their confusion to convert this opportunity into actual learning gains. Learning occurs when learners engage in cognitive activities beneficial to learning (e.g., reflection, deliberation, problem solving) during the process of confusion resolution. Unfortunately, learners are not always able to resolve their confusion on their own. The inability to resolve confusion can be due to a lack of knowledge, motivation, or skills. The present dissertation explored methods to aid confusion resolution and ultimately promote learning through a multi-pronged approach. First, a survey revealed that learners prefer more information and feedback when confused and that they preferred different interventions for confusion compared to boredom and frustration. Second, expert human tutors were found to most frequently handle learner confusion by providing direct instruction and responded differently to learner confusion compared to anxiety, frustration, and happiness. Finally, two experiments were conducted to test the effectiveness of pedagogical and motivational confusion regulation interventions. Both types of interventions were investigated within a learning environment that experimentally induced confusion via the presentation of contradictory information by two animated agents (tutor and peer student agents). Results showed across both studies that learner effort during the confusion regulation task impacted confusion resolution and that learning occurred when the intervention provided the opportunity for learners to stop, think, and deliberate about the concept being discussed. Implications for building more effective affect-sensitive learning environments are discussed.

Table of Contents

Chapter		Page
1.	Introduction	1
	Affect-Sensitive Intelligent Tutoring Systems	4
	Confusion and Learning	5
	Confusion Induction Learning Environments	10
	Present Research Objectives	13
2.	Study 1: Learner Preferences for Confusion Regulation Interventions	14
	Survey and Data Collection	15
	Results & Discussion	17
3.	Study 2: Expert Human Tutor Responses to Learner Confusion	20
	Expert Tutoring Corpus	20
	Data Treatment	22
	Results & Discussion	23
4.	Study 3: Pedagogical Confusion Interventions	43
	Learning Activity	43
	Method	44
	Results & Discussion	57
5.	Study 4: Motivational Confusion Interventions	86
	Method	86
	Results & Discussion	89
6.	General Discussion	110
	Overview of Research	110
	Limitations and Future Directions	118
	Conclusion	121
References		123
Appendices		
A.	Survey Research Questions (Study 1)	136
B.	Informed Consent Form (Study 1)	138
C.	Data Release Agreement (Study 1)	139
D.	Debriefing Form (Study 1)	140
E.	Academic Grit Scale	141
F.	School Failure Tolerance Scale	142
G.	Motivated Strategies for Learning Questionnaire	144
H.	Attributional Complexity Scale	146
I.	Demographics Questionnaire	148
J.	Informed Consent Form (Studies 3 & 4)	149

K. Data Release Agreement Form (Studies 3 & 4) 150
L. Debriefing Form (Study 3) 151
M. Debriefing Form (Study 4) 152
N. Flaw-Identification Task Pretest (Studies 3 & 4) 153
O. Near Transfer Task Posttest (Studies 3 & 4) 155
P. Far Transfer Task Posttest (Studies 3 & 4) 158
Q. Design-A-Study Task Posttest (Studies 3 & 4) 160
R. Zoner & Nemet (2002) Coding Scheme 162
S. IRB Approval 163
T. Induction × Confusion × Intervention × Regulation Effort Interaction Results from Study 3 164
U. Induction Condition Differences for the Induction × Confusion × Intervention × Regulation Effort for the Far Transfer Task in Study 3 167
V. Induction Condition Differences for the Induction × Confusion × Intervention × Regulation Interaction for the Near Transfer and Design-A-Study Tasks in Study 4 168

List of Tables

Table		Page
1.	Mean (SD) of learner preferences for interventions	18
2.	Mean (SD) of dialogue moves following instances of confusion overall	25
3.	Mean (SD) of feedback dialogue moves following instances of confusion types	36
4.	Mean (SD) of motivational dialogue moves following instances of confusion types	37
5.	Mean (SD) of tutor question dialogue moves following instances of confusion types	40
6.	Mean (SD) of tutor instruction dialogue moves following instances of confusion types	41
7.	Excerpt of trialogue of induction and post-intervention phases from *True-False* condition	51
8.	Excerpt of trialogue of intervention phase from *True-False* condition	53
9.	Proportional occurrence of induction phase dependent measures	59
10.	Proportional occurrence of confusion regulation process dependent measures	62
11.	Proportional occurrence of argument quality dependent measures	68
12.	Proportional occurrence of post-intervention phase dependent measures	71
13.	Proportional occurrence for learning outcome dependent measures	77
14.	Proportional occurrence of transfer task performance	85
15.	Excerpt of trialogue of intervention phase from *True-False* condition	88
16.	Proportional occurrence of induction phase dependent measures	91
17.	Mean (SD) of classification performance across groups	93
18.	Proportional occurrence of argument quality dependent measures	97
19.	Proportional occurrence of post-regulation phase dependent measures	99

20. Proportional occurrence of learning measures 102

21. Proportional occurrence of near transfer and design-a-study tasks 105

22. Summary of main findings from Studies 3 and 4 115

Interventions to Regulate Confusion during Learning

Intelligent tutoring systems (ITS) are designed to increase learning through adaptive, individualized instruction and scaffolding. Adaptive responses can range from feedback about the quality of a learner's response (e.g., "Yes, that's correct.") to mini-conversations that break down the main problem down into smaller sub-problems (e.g., scaffolding). ITSs that respond to the cognitive states of learners (i.e., response quality) have been found to be similar in effectiveness to human tutors (VanLehn, 2011). Recently, ITSs have been developed that respond to both the cognitive and affective states of learners (Arroyo et al., 2009; Burleson & Picard, 2007; Chaffar, Derbali, & Frasson, 2009; Conati & Maclaren, 2009; D'Mello, Craig, Fike, & Graesser, 2009; D'Mello & Graesser, 2012a; Forbes-Riley & Litman, 2009; Robison, McQuiggan, & Lester, 2009). Although these affect-sensitive ITSs are also effective, there is still an open question as to what the best method is to respond to learner affect to promote engagement and learning.

ITS responses to learner cognitive states attempt to correct erroneous beliefs and increase understanding of a concept. ITS responses to learner affective states, on the other hand, attempt to manage learner affect to maintain or return to a state that is conducive for learning. Engagement, for example, is a state that has been found to be positively correlated with learning (Craig, Graesser, Sullins, & Gholson, 2004; Graesser & D'Mello, 2012; Graesser et al., 2008). Thus, an affective-sensitive ITS could have the goal to keep learners in a state of engagement or return learners to a state of engagement when other affective states occur. In fact, when affective state transitions were investigated in AutoTutor, a natural language mixed-initiative ITS (Graesser et al., 2004),

a cycle of affect transitions was identified that consisted of oscillations between engagement and confusion (see Figure 1, D'Mello & Graesser, 2012b). This productive cycle involved learners detecting impasses (engagement → confusion) and then resolving those impasses (confusion → engagement). However, another cycle was also identified that consisted of transitions from confusion to frustration and then oscillations between frustration and boredom. Based on these findings, an affect-sensitive ITS with the goal to keep learners in a state of engagement could then focus on instances of confusion. The ITS could intervene when confusion occurs to help learners transition to the productive state of engagement as opposed to the unproductive state of frustration. This is one example of how an affect-sensitive ITS could operate; however, there is not a consensus as to which affective state is best for learning and what type of intervention is most effective to achieve this goal for particular groups of learners.

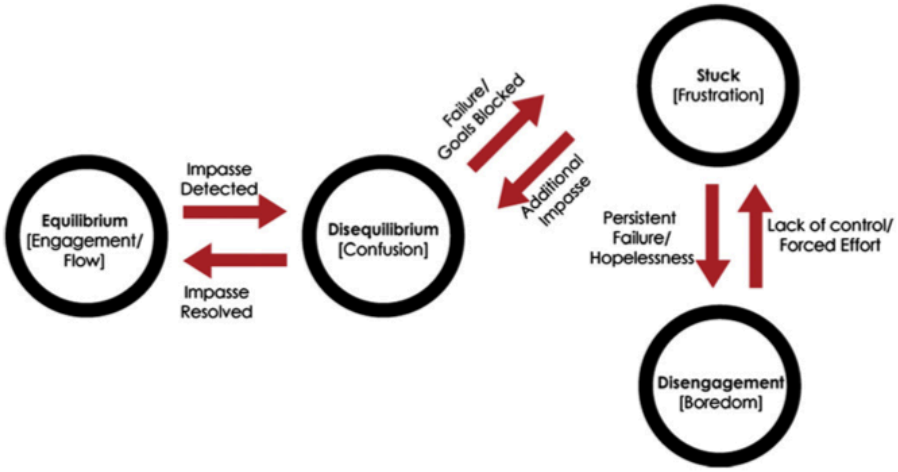

Figure 1. Observed Model of Affect Dynamics, reprinted from D'Mello & Graesser (2012b).

Affect-sensitive ITSs must first determine which affective states will receive adaptive responses. Research on shorter learning sessions (i.e., 30 min to 1.5 hr) have found a set of learning-centered affective states that frequently occur in these learning contexts (Arroyo et al., 2009; Baker, D'Mello, Rodrigo, & Graesser, 2010; Burleson & Picard, 2007; Chaffar et al., 2009; Conati & Maclaren, 2009; D'Mello, 2013; D'Mello et al., 2009; Forbes-Riley & Litman, 2011; Graesser & D'Mello, 2012; Lehman, Matthew, D'Mello, & Person, 2008; Robison et al., 2009; Rodrigo & Baker, 2011b). The learning-centered affective states (i.e., anxiety, boredom, confusion, curiosity, engagement, delight, frustration, surprise) (Calvo & D'Mello, 2011; D'Mello, 2013; Rodrigo & Baker, 2011a) can be contrasted with the universal, life experience set of basic emotions (i.e., anger, contempt, disgust, fear, happy, sad, surprise) (Ekman, Friesen, & Ellsworth, 1972) and the academic achievement emotions, which occur over longer time periods such as an academic semester or year (Pekrun, 2010). These achievement emotions are associated with (a) outcomes (*achievement*, e.g., contentment, anxiety, and frustration), (b) different topics (*topic*, e.g., empathy for the protagonist in a novel), (c) interpersonal interactions (*social*, e.g., pride, shame, and jealousy), and (d) new information (*epistemic*, e.g., surprise and confusion). Interactions with ITSs generally occur in the time frame of shorter learning sessions and are focused on learning. Thus, the learning-centered affective states seem to be the affective states to which ITSs should provide adaptive responses. However, there is still the issue of how to respond to these affective states to promote learning. Two ITSs that respond to learner affective states are discussed next.

Affect-Sensitive Intelligent Tutoring Systems

Crystal Island is a narrative-centered ITS that immerses learners in a virtual world to learn microbiology while solving a medical mystery (McQuiggan, Mott, & Lester, 2008; McQuiggan, Robison, & Lester, 2010). Empathetic agents then adaptively responded to learners' current affective states in one version of Crystal Island (Robison et al., 2009). The agents responded either in parallel or reactively. Parallel responses involved agents mirroring learners' affective states (e.g., learner is frustrated and the agent displays frustration), whereas reactive responses involved agents displaying the desired affective state for learners (e.g., learner is frustrated and the agent displays empathy). Agent emotions were displayed through text-based dialogue. This intervention was found to influence transitions between affective states. However, this intervention was not effective for all affective states, particularly confusion.

Affective AutoTutor is a version of the previously mentioned AutoTutor that detects and responds to learner affective states (D'Mello, Lehman, & Graesser, 2011). Responses to learner affective states involved three components: (1) content of agent utterance, (2) agent facial expression, and (3) agent speech. The content of the agent's utterance was an affective statement based on attribution theory (Batson, Turk, Shaw, & Klein, 1995; Heider, 1958; Weiner, 1986), cognitive disequilibrium theory (Festinger, 1957; Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005; Piaget, 1952), and pedagogical experts. The affective statements consisted of an empathetic or motivational response and attributed the learner's emotion to either the material or the tutor. When confusion was detected, for example, AutoTutor would say, "Some of this material can be confusing. Just keep going and I'm sure you'll get it." Results showed that learning gains were

greater for the Affective AutoTutor than the non-affective version, but only for a subset of learners. Specifically, Affective AutoTutor was most effective for low prior knowledge learners.

The interventions used in both Crystal Island and Affective AutoTutor attempted to apply the same type of intervention to multiple affective states, although responses were tailored to the individual affective state (e.g., boredom versus confusion). It may be the case, however, that different types of interventions are needed for different affective states. In other words, resolving confusion and overcoming boredom may not be achieved by the same type of intervention. When developing interventions for specific affective states, it would be advantageous to begin with an affective state that has the potential to greatly impact learning. Confusion is one such affective state. Similar to boredom and frustration, confusion is a negatively-valenced affective state. However, unlike boredom and frustration, confusion has been found to be positively correlated with learning (Craig et al., 2004; Graesser & D'Mello, 2012; Graesser et al., 2008). As previously mentioned, confusion is involved in both the productive (engagement) and unproductive (frustration, boredom) affective cycles. Confusion then has the potential to positively impact learning if successfully resolved, but also the potential to lead to frustration and disengagement if not properly addressed. Next, the characteristics of confusion during learning are discussed.

Confusion and Learning

Confusion is an epistemic or knowledge affective state (Pekrun & Stephens, 2012; Silvia, 2010) that is triggered by anomalies, breakdowns, contradictions, impasses, and uncertainty about how to proceed (Brown & VanLehn, 1980; Caroll & Kay, 1988;

D'Mello, Lehman, Pekrun, & Graesser, 2014; VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003). In short, confusion indicates that there is a problem with the current state of one's knowledge (Piaget, 1952). Confusion creates an opportunity for learning because the problematic aspect of the learner's knowledge has been highlighted and can then be addressed and corrected. Cognitive activities that are beneficial for learning (e.g., reflection, deliberation, problem solving) can be triggered by experiences of confusion in an effort to correct the problem in one's knowledge. Confusion resolution occurs when the problem in one's knowledge has been corrected. It is this process of confusion resolution that ultimately leads to learning, which is consistent with cognitive disequilibrium (Festinger, 1957; Graesser et al., 2005; Piaget, 1952) and impasse-driven theories of learning (Brown & VanLehn, 1980; VanLehn et al., 2003). This type of productive confusion should be contrasted with hopeless or unresolved confusion (D'Mello & Graesser, 2012b). Hopeless confusion is an instance of confusion that the learner cannot resolve either due to a lack of motivation or knowledge. These two types of confusion can be seen in the affective transitions identified by D'Mello and Graesser (2012b) (see Figure 1). These two types of confusion can be viewed as virtuous and pathological outcomes. The mechanism that discriminates between the virtuous and pathological outcomes of confusion is whether or not cognitive disequilibrium is resolved. Thus, the mechanism-based intervention target becomes cognitive disequilibrium or confusion resolution. Hopeless confusion could then be overcome if an appropriate intervention is deployed to create opportunities for and promote confusion resolution. For example, supports through more information and motivation as well as scaffolding could help learners to persist through these struggles and clear up any

uncertainties. Thus, it is important to determine which types of interventions will help learners resolve their confusion when they are unable to resolve it on their own.

An intervention that places the responsibility of confusion resolution on the learner is consistent with Piagetian theory (1952). Piagetian theory suggests that learners must experience cognitive conflict for a sufficient amount of time before they adequately deliberate and reflect via self-regulation. Based on the findings from over 100 hours of one-to-one human tutoring, VanLehn et al. (2003) recommended a similar response to learner impasses that involves three steps: (1) prompting the learner to reason and arrive at a solution, (2) prompting the learner to explain their solution, and (3) providing the solution with an explanation only if the learner fails to arrive at an answer. This research suggests that learners can resolve their confusion, created by an impasse, through continued reflection, deliberation, and problem solving.

Vygotskian theory (1978), on the other hand, suggests a more guided, direct approach to addressing confusion solution. This perspective would suggest that it is not productive to have struggling learners spend too much time in a state of confusion and that ITSs should provide direct questions and explanations to aid confusion resolution. This type of intervention could also be manifested in breaking down the original problem into smaller, more manageable sub-problems for the learner. This strategy was found in a corpus of 50 hr of one-to-one expert human tutoring sessions (Lehman et al., 2008). Tutors were found to break down the original problem after instances of learner confusion. A similar pattern was also derived from another corpus of one-to-one human tutoring that investigated responses to learner uncertainty (Forbes-Riley & Litman, 2007). The response involved three components: (1) drawing the learner's attention to the

impasse, (2) providing additional hints and questions to the learner, and (3) providing the learner with additional information that is needed to resolve the impasse. The pattern found by Forbes-Riley and Litman was then used to create an uncertainty-adaptive version of ITSpoke (Forbes-Riley & Litman, 2011), which is discussed later.

Another type of intervention can be derived from attribution theory (Batson et al., 1995; Heider, 1958; Weiner, 1986), which is similar to the approach used in the previously described Affective AutoTutor. Attribution theory suggests that perceived causes of outcomes (success, failure) will impact future behaviors. Two learners who experience confusion can ascribe different causal attributions and this attribution will impact whether a learner engages in effortful cognitive activities to resolve their confusion or disengages from the task. Attributions vary on three factors: locus (internal, external), stability (stable, unstable), and control (controllable, uncontrollable) (Weiner, 2010). When learners attribute the cause to their ability, for example, the attribution is generally internal, stable, and uncontrollable; whereas an attribution to effort would be internal, unstable, and controllable. Attributions to effort when failure occurs have been found to be more beneficial than to ability in the case of academic achievement because learners have the ability to change the amount of effort exerted in the future.

Misattribution training is one approach to alter learners' attributions to encourage persistence after failure (Reisenzein, 1983; Schachter & Singer, 1962). A self-attribution of low ability after failure can cause a reduction in self-esteem and cause learners to disengage from the task. However, an external attribution can avoid this threat to self-esteem through removing the control or responsibility of failure from the learner. Affective AutoTutor utilizes this type of approach by placing the responsibility of

confusion on the tutor agent or to the difficulty of the material when confusion occurs. In other words, AutoTutor places the responsibility for confusion on either the difficulty of the material or the prior explanations delivered by AutoTutor.

Previous research has also utilized attribution retraining to shift learners from ability to effort (Anderson, 1983; Andrews & Debus, 1978; Medway & Venino, 1982; Zoeller, Mahoney, & Weiner, 1983), stable to unstable (Wilson & Linville, 1982; 1985), and uncontrollable to controllable causes for failure (Perry, Stupinsky, Hall, Chipperfield, & Weiner, 2010). The attribution retraining to shift learners from ability attributions to effort attributions is also very similar to approach adopted by Dweck (1999) that encourages learners to adopt an incremental as opposed to fixed mindset. This attribution retraining has been found to positively impact learning in both the short-term and long-term (e.g., next semester) by encouraging learners to persist after failure (Perry et al., 2010; Wilson & Linville, 1982, 1985). This approach can be applied to experiences of confusion, which are not the same as failure but do frequently involve a negative attribution and require persistence and resilience to reach a successful resolution.

UNC-ITSpoke is the uncertainty adaptive version of ITSpoke, which is a spoken-dialogue ITS that tutors qualitative physics (Litman & Forbes-Riley, 2006). Adaptive responding was based on uncertainty (present, absent) and the quality of learner responses (correct, incorrect). The combination of uncertainty and response quality resulted in four outcomes that differed based on the severity of the impasse experienced by the learner: (1) correct + certain (least severe), (2) correct + uncertain, (3) incorrect + uncertain, and (4) incorrect + certain (most severe). Greater learning gains were found for learners who interacted with UNC-ITSpoke than those who interacted with the regular

version of ITSpoke (Forbes-Riley & Litman, 2011). However, this pattern was only found for those learners who experienced more uncertainty and received more support from UNC-ITSpoke. This limitation raises an interesting issue about evaluating the effectiveness of uncertainty and confusion regulation interventions.

The previously discussed affect-sensitive ITSs take a reactive approach by responding after a particular affective state has naturally occurred. Affect-sensitive ITSs could take a more proactive approach, however, by creating learning opportunities through the induction of confusion in addition to reactions to confusion when affective states naturally occur. This type of affect-sensitive ITS would aid learners by providing support for confusion resolution, but also by creating opportunities for learners to reach a deeper understanding. Three methods of confusion induction and their impact on learning are discussed next.

Confusion Induction Learning Environments

In the first example, D'Mello and Graesser (in press) investigated the use of device breakdowns as a method of confusion induction in two experiments. Participants read illustrated texts of everyday devices and were then presented with the same illustrated text plus an additional breakdown prompt (Breakdown Condition). The cylinder lock, for example, had the following breakdown prompt: "A person puts the key into the lock and turns the lock but the bolt doesn't move." Participants were then asked to determine why the device was not functioning. In the control condition participants either re-read the illustrated text (Experiment 1) or re-read the illustrated text with instructions to focus on a key part of the device (Experiment 2). More confusion was reported by participants when in the breakdown condition compared to the control

condition in both experiments. Confusion over time was also investigated by having participants complete a continuous rating of their confusion level via a retrospective confusion judgment. Analyses revealed the occurrence of two main patterns of confusion: partially-resolved and unresolved. Participants who were able to partially resolve their confusion outperformed their counterparts who were unable to resolve their confusion on a device comprehension task.

In the second example, three experiments investigated the presentation of contradictory information as a method of confusion induction (D'Mello et al., 2014; Lehman et al., 2013). Contradictory information was presented by animated pedagogical agents in each of these experiments. Two agents (tutor and peer student agents) presented their opinions while discussing the scientific merits of research case studies. For example, one case study described a miraculous new diet pill that caused significant weight loss in just one month. For this case study, the two agents and the human learner discussed whether or not the control group used in this study was appropriate. Contradictory information was presented via the agents' opinions. Agents could agree and present correct opinions (True-True), agree and present incorrect opinions (False-False), or disagree with each other (True-False, False-True). In the True-False condition the tutor agent presented a correct opinion and the student agent disagreed and with an incorrect opinion, whereas it was the tutor agent who disagreed with an incorrect opinion and the student agent who presented a correct opinion in the False-True condition. The human learner was then invited to provide his or her opinion after the agents had each presented their opinion. Confusion was successfully induced when contradictory information conditions were compared to the no-contradiction control condition (True-True). In

addition, learners who were successfully confused by the contradictory information performed better on learning measures (Experiments 1-3), including a difficult transfer of knowledge task (Experiment 3).

In the third example, false feedback was investigated as a method of confusion induction (Lehman, D'Mello, & Graesser, in preparation). Learners diagnosed flaws in research case studies with the guidance of an animated pedagogical tutor agent. After the learner diagnosed the flaw in the case study via a forced-choice question, the tutor agent then provided feedback on the quality of the diagnosis. The feedback could either be accurate or inaccurate, regardless of actual response quality (correct, incorrect). Learners reported more confusion when they were correct and received inaccurate, negative feedback than when they received accurate, positive feedback. Learners performed best on a difficult far transfer task when they received false feedback (positive or negative) and were successfully confused based on on-line judgments than when they received accurate feedback (positive or negative).

Although learning gains associated with confusion induction were found for all three methods of confusion induction, it was not the case that all learners were able to successfully resolve their confusion and learn the material. This was likely due to the fact that most of these experiments did not provide any aid or scaffolding for confusion resolution. Two experiments provided learners with an explanatory text to aid confusion resolution (contradictory information: Experiment 3; false feedback experiment). The explanatory text, however, may not have been sufficient to aid all learners in the effort to resolve their confusion. This dissertation addresses this issue by exploring interventions to regulate confusion during learning.

Present Research Objectives

The present dissertation adopts a multi-prong approach to better understand what interventions are effective to regulate confusion and promote learning. The remainder of the proposal is organized into five sections. First, learner preferences for affective interventions were investigated in a survey study. Second, the way in which expert human tutors respond to learner confusion were investigated. In sections three and four, interventions to regulate confusion were investigated within a learning environment that experimentally induces confusion via the presentation of contradictory information from two agents. Learners engaged in a trialogue (three-party conversation) with two animated pedagogical agents (tutor and peer student agents) to evaluate the scientific merits of research case studies. Interventions built on previous research by exploring two types of interventions. Pedagogical and motivational interventions were investigated in two experiments. The pedagogical interventions encouraged learners to engage in the cognitive activities needed for confusion resolution (e.g., reflection, deliberation). The motivational interventions were designed to motivate learners to persist through experiences of confusion and to put in effort to resolve their confusion through changes in their causal attributions. Finally in section five, the findings from all four studies, limitations in the present research, and future work are discussed in a general discussion.

2. Study 1: Learner Preferences for Confusion Regulation Interventions

Recent research has explored various theoretically- or empirically-derived interventions for adaptively responding to affective states. However, there has been a paucity of research exploring learner preferences for interventions. One exception is a survey study that explored learner methods of self-regulation for affect while studying academic material (Strain, Gross, & D'Mello, 2012). This study reported that some self-regulation strategies were viewed as more effective than others. In particular, learners reported that quiet seeking, taking a break, positive rumination, engaging in a learning strategy (e.g., taking notes, highlighting), and making a game out of studying were helpful strategies. The current dissertation further explored learner preferences for affect interventions in line with this previous study.

It is possible, however, that learner preferences may not be informative due to a lack of meta-affect knowledge. This possibility stems from the extensive previous research that has shown learners have poor metacognitive knowledge (Dunlosky & Lipko, 2007; Glenberg & Epstein, 1985; Graesser, D'Mello, & Person, 2009; Hacker, Dunlosky, & Graesser, 2009). In general, meta-comprehension ratings have been found to correlate with actual learner performance very poorly (Dunlosky & Lipko, 2007; Glenberg, Wilkinson, & Epstein, 1982; Maki, 1998). Learner awareness of affect and its relation to current understanding could be similar to metacognitive awareness. If this is the case, then learner preferences for confusion regulation interventions may not be particularly informative for developing effective interventions. However, there is an alternative position that affect is more salient for learners (Damasio, 1994; Izard, 1993;

Mayer, Salovey, & Caruso, 2004; Scherer, 1993) and thus learner reflections could be insightful.

Research in cognitive neuroscience has suggested that there are separate 'hot' and 'cold' paths for affect and cognition, respectively (Botvinick, Cohen, & Carter, 2004; Bush, Luu, & Posner, 2000; D'Esposito, Postle, & Rypma, 2000; Ochsner & Gross, 2005). In particular, increased activity in the anterior cingulate cortex has been found when there is conflict in information processing and negative outcomes in order to reduce conflict and negative outcomes in the future through the action selection process (Botvinick et al., 2004).

Survey and Data Collection

Participants. Participants were 105 undergraduate students and received course credit for participation. Participants ranged in age from 18 to 56 years old (M = 21.4, SD = 6.66). There were 97 females and 8 males. Thirty-two percent of participants were African American, 53% were Caucasian, 8% were Asian, 1% were Hispanic, and 6% were other.

Survey. The survey consisted of two phases: preferences on interventions for emotion regulation during learning (see Appendix A) and individual difference measures (see Appendices E-I). To determine preferences for interventions for confusion regulation, participants were asked a series of self-report questions. Specifically, participants were asked: *When you are CONFUSED during learning, how helpful would you find each of the following for overcoming your CONFUSION?* Learning was defined for participants as any experience in a classroom, working alone, or working with a tutor in which you are attempting to learn some material (see Figure 2).

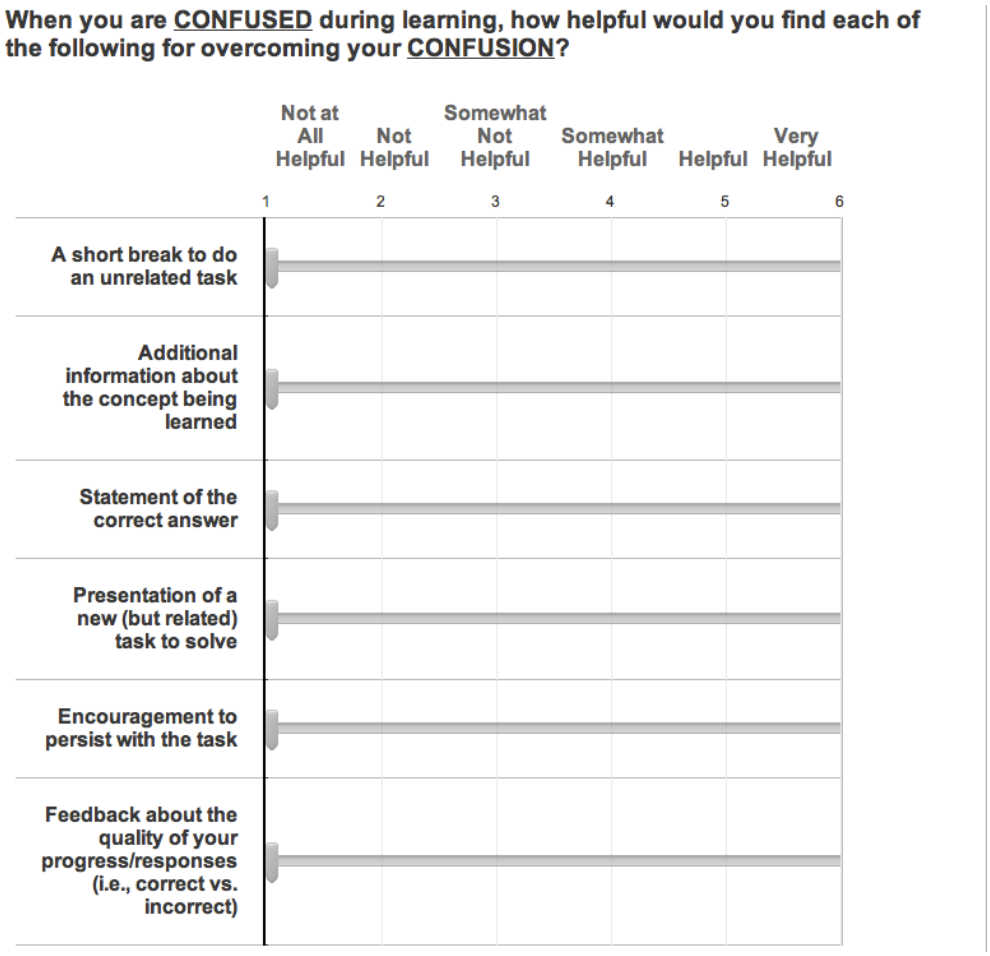

Figure 2. Screenshot of Confusion Intervention Survey Questions.

Participants then rated each potential intervention on a 6-point scale with the following anchors: 1 – Not at All Helpful and 6 – Very helpful, with all values anchored (see Figure 2). The presentation of each response option was randomized for each participant. The six confusion regulation interventions were: (1) additional information about the concept being learned, (2) encouragement to persist with the task, (3) presentation of a new (but related) task to solve, (4) feedback about the quality of your responses (i.e., correct vs. incorrect), (5) correct answer, and (6) a short break to do an unrelated task. Participants repeated this process for experiences of boredom and frustration as well.

Next, participants completed several individual differences measures. Participants completed the Academic Grit Scale (AGS, Duckworth, Peterson, Matthews, & Kelly, 2007, see Appendix E), School Failure Tolerance Scale (SFT, Clifford, 1984, see Appendix F), Motivated Strategies for Learning Questionnaire (MSLQ, Pintrich & DeGroot, 1990, see Appendix G), Attributional Complexity Scale (ACS, Fletcher, Fernandez, Peterson, & Reeder, 1986, see Appendix H), and a demographics questionnaire (i.e., gender, ACT score, year in school, etc., see Appendix I). These measures were selected because they assess preferences for challenging material and responses to academic challenges like those posed by confusion experiences.

Results and Discussion

The analyses were conducted in two phases. First, there was a comparison of each type of intervention for confusion. This analysis was conducted to determine which type of intervention participants most preferred when confused during learning. Second, there was a comparison of each type of intervention for each affective state (boredom, confusion, frustration). This analysis was conducted to determine which type of intervention was viewed as uniquely effective for a specific affective state or was beneficial to multiple affective states. The analyses were conducted with ANOVAs that had intervention or intervention × affective state as the factors.

Within Confusion Comparison. A Repeated Measures ANOVA investigating learner preferences for confusion regulation interventions was found to be significant, $F(5,105) = 29.6$, $p < .001$, $Mse = 1.35$, $\eta^2 = .218$. Table 1 shows learner ratings for preference of each method of intervention. Post hoc analyses (least significant difference) were conducted to investigate which intervention was preferred by learners. The general

pattern for preference was: More Information > (Feedback = Correct Answer) > Encouragement > (New Task = Break). The only exception to this general pattern was that learners equally preferred being provided with the correct answer and encouragement.

Learners appear to generally prefer interventions that involved staying on task when confused. Specifically, learners felt that receiving more information, feedback, and the correct answer would all help them overcome their confusion. In contrast, changing the task (new related task or break) was least preferred. Encouragement was found to be less preferred to the stay on task interventions, but more preferred than the change task interventions.

Table 1
Mean (SD) of Learner Preferences for Interventions

	Boredom		**Confusion**		**Frustration**	
	M	*SD*	*M*	*SD*	*M*	*SD*
More Information	3.50	1.67	5.15	1.16	4.94	1.32
Feedback	3.77	1.70	4.88	1.34	4.86	1.34
Correct Answer	3.05	1.59	4.82	1.17	4.62	1.41
Encouragement	3.98	1.66	4.64	1.41	4.67	1.45
Take a Break	4.79	1.61	3.64	1.75	4.71	1.49
New Related Task	4.55	1.51	3.85	1.52	4.13	1.66

Across Emotion Comparison. Next, a Repeated Measures ANOVA was conducted to investigate the emotion (boredom, confusion, frustration) × intervention interaction (see Table 1). The interaction was significant $F(10,105) = 37.0$, $p < .001$, $Mse = 1.15$, $\eta^2 = .259$. Post hoc analyses revealed that there were differences in learner preferences for interventions based on the affective state being addressed. More information and provide correct answer as interventions were most preferred for

occurrences of confusion, followed by frustration, and least preferred for boredom (Confusion > Frustration > Boredom). Encouragement and feedback were equally preferred for confusion and frustration, but least preferred for boredom (Confusion = Frustration > Boredom). Shifting to a new but related task was most preferred for instances of boredom, followed by frustration, and least preferred for confusion (Boredom > Frustration > Confusion). Finally, taking a break was equally preferred for boredom and frustration, but least preferred for confusion (Boredom = Frustration > Confusion).

Confusion differed from both boredom and frustration in terms of the interventions that learners felt would be helpful. However, it was the case that confusion was similar in some respects to frustration. In particular, feedback and encouragement were viewed as equally helpful for overcoming confusion and frustration. This similarity could be due to the fact that learners may perceive instances of confusion and frustration as similar. In fact it has been proposed and found in interactions with AutoTutor that when learners stay in confusion for too long and are unable to resolve their confusion, they can transition into frustration (D'Mello & Graesser, 2012b). Thus, it may be that some learners were referencing hopeless or unresolvable confusion when rating which intervention would be most helpful. In future studies, it would be helpful to determine how learners perceive confusion and frustration and what type of confusion and frustration experience they are referencing when rating potential interventions.

3. Study 2: Expert Human Tutor Responses to Learner Confusion

Expert Tutoring Corpus

The corpus consisted of 50 tutoring sessions between ten expert tutors and 39 learners (Cade, Copeland, Person, & D'Mello, 2008; Person, Lehman, & Ozbun, 2007). Expert status was defined as: licensed to teach at the secondary level, five or more years of ongoing tutoring experience, employed by a professional tutoring agency, and highly recommended by local school personnel. The learners were all having difficulty in a science or math course and were either recommended for tutoring by school personnel or voluntarily sought professional tutoring help. All learner and tutor pairs were working together prior to this study. Some learner and tutor pairs were recorded for two tutoring sessions. Therefore, the unit of analysis was the tutor-learner dyad. The subjects studied were algebra (28%), basic math (2%), biology (20%), chemistry (8%), geometry (26%), physics (8%), and standardized test preparation (8%).

Fifty-five percent of learners were female and 45% were male. Sixty-nine percent of learners were Caucasian, 28% were African American, and 3% were Asian American. Learners varied in age from 13 to 25 years old with a median age of 16 years old, ranging from middle school to an adult returning to get her general education diploma. Of the 39 learners, four were home schooled, 17 attended public schools, and 18 attended private schools. Tutors reported the socio-economic status of their learners as 12 upper class, 12 upper-middle class, and 15 middle class.

Each session lasted approximately 1 hr. All sessions were videotaped with a camera and positioned at a great enough distance to not disturb the tutoring session but close enough to record audio and visual data. The researcher left the room during the

tutoring session. The videos were digitized and then transcribed. Transcripts were then coded with respect to tutor dialogue moves, learner dialogue moves, and learner affective states (see below).

Tutor Dialogue Moves Coding Scheme. The 24-item tutor dialogue move coding scheme (Person et al., 2007) was divided into groups based on similar functions within the tutoring session: *direct instruction* (example, counterexample, preview, summary, provide correct answer, direct instruction), *question* (new problem, simplified problem, prompt, pump, hint, forced-choice), *feedback* (positive, neutral, negative), *motivational statement* (humor, attribution, general motivation, solidarity), *conversational "Okay" (i.e., backchannel feedback),* and *off-topic*.

Learner Dialogue Moves Coding Scheme. The 16-item learner dialogue coding scheme was divided into eight groups based on the function of each move: *answer* (correct, partially-correct, vague, error-ridden, none), *question* (common ground, knowledge deficit), *misconception* (e.g., "I thought it was the other way around"), *metacomment* (e.g., "I don't know",), *work-related action* (think aloud, read aloud, work silently), *socially motivated action* (social coordination, acknowledge), *gripe*, and *off-topic*.

Learner Affective State Coding Scheme. The 12-item learner affective state coding scheme (Lehman et al., 2008) consisted of both learning centered affective states (anxiety, confusion, curiosity, eureka, frustration) and Ekman's basic emotions (anger, contempt, disgust, fear, happy, sad, surprise). The four most prominently occurring affective states during the tutoring sessions were anxiety, confusion, frustration, and happiness. They accounted for 93.0% of the affective states that learners experienced

during the expert tutoring sessions. The present proposal will then focus on these four affective states for the proposed analyses.

Data Treatment

A previous analysis investigated tutor responses to learner affective states (Lehman et al., 2008). However, this previous analysis had two limitations. First, the analysis only took into consideration a single tutor dialogue move in response to an affective state. It is possible that tutor responses to affective states involve multi-move or even multi-turn responses. For example, both VanLehn et al. (2003) and Forbes-Riley and Litman (2007) reported that human tutor responses to impasses and uncertainty involved multi-turn interactions between the tutor and the learner. Second, the prior analysis considered all instances of an emotion to be equivalent. It is possible, for example, that tutors respond differently to a learner who is confused and asks a question and a learner who is confused and responds incorrectly. The analyses attempt to address these two limitations.

Tutor responses to confusion were defined as 20 dialogue turns after the confusion experience. Twenty dialogue turns were selected as the unit of analysis to capture approximately 10 tutor and 10 learner dialogue turns following each instance of confusion. In other words, when an instance of learner confusion was identified in dialogue turn 5, for example, dialogue turns 6 through 26 were included in the present analyses. To conduct the analyses, each instance of confusion was identified in the expert tutoring corpus and the 20 subsequent dialogue turns were extracted. In addition, the learner dialogue move associated with the instance of confusion was also recorded. This was done to determine if tutor responses differ based on the combination of learner

affective and cognitive states. This process was also completed for instances of anxiety, frustration, and happiness in order to compare tutor responses to the four most frequently occurring affective states in the expert tutoring corpus. Finally, tutor responses to learner answers (correct, partially-correct, vague, error-ridden, and none) were also collected to determine if tutor responses differed based on the presence or absence of confusion.

Results and Discussion

The analyses were conducted in two phases. First, occurrences of tutor dialogue moves were investigated for instances of confusion in general. Second, occurrences of dialogue moves were investigated based on confusion and the co-occurring cognitive state. Cognitive states were bounded by associated dialogue moves pertaining to answer types (correct, partially-correct, vague, error-ridden, none), questions (common ground, knowledge deficit), and metacognitive statements (metacomment, misconception). In previous research on this expert tutoring corpus, confusion was found to significantly co-occur with all types of learner incorrect responses, questions, and metacomments, but not misconceptions (Lehman, D'Mello, & Person, 2010). Hypotheses based on impasse-driven theories of learning (Brown & VanLehn, 1980; VanLehn et al., 2003) and cognitive disequilibrium theory (Festinger, 1957; Graesser et al., 2005; Piaget, 1952) would expect the tutors to encourage learners to resolve their confusion on their own with only vague questions (e.g., pumps, hints) to trigger the learner to provide a response and explanation for their response. In contrast, a hypothesis based on Vygotskian theory (1978) would expect the tutors to break the original problem down into smaller sub-problems and engage in more directly guided scaffolding. Finally, a hypothesis based on the INSPIRE Model (Lepper & Woolverton, 2002) would expect the tutors to provide

supportive, motivational statements or be nurturing to the learners. The analyses allowed for the investigation of each hypothesis.

The analyses included ANOVAs and paired-samples *t*-tests. When *t*-tests were computed instead of ANOVAs it was due to the fact that every dialogue move (e.g., learner questions, learner answers), affective state, or confusion-cognitive state pair (e.g., confusion-vague answer) did not occur in every expert tutoring session. In order to maintain a large enough N to conduct meaningful analyses *t*-tests were necessary.

Instances of Confusion Overall. Instances of confusion overall were investigated in five contexts. First, the pattern of tutor dialogue moves following instances of confusion was investigated (confusion only context). Next, the occurrence of dialogue moves following instances of confusion were compared to dialogue move occurrences in other contexts in the expert tutoring session: (2) overall, (3) other learner affective states (anxiety, frustration, happiness), (4) learner questions, and (5) learner answers.

Within confusion comparison. A Repeated Measures ANOVA was conducted to investigate the occurrence of tutor dialogue moves following instances of confusion overall. Table 2 shows the proportional occurrence of all possible tutor dialogue moves. Note that the dialogue moves in Table 2 do not add up to 1 because both tutor and learner dialogue moves occurred, however, only tutor dialogue moves are included in the present analyses. The ANOVA was significant, $F(23, 49) = 139$, $p < .001$, $\eta^2 = .743$. The following overall pattern was found when Bonferroni post hoc analyses were conducted: direct instruction > (positive feedback = off-topic = conversational ok) > (simplified problem = comprehension gauging question = prompt = repetition = hint = new problem = provide correct answer = negative feedback = neutral feedback = humor = attributional

acknowledgement = example) > (pump = general motivational statement = summary = paraphrase = forced-choice question = preview = solidary statement = counter example).

Table 2
Mean (SD) of Dialogue Moves following Instances of Confusion Overall

Tutor Dialogue Moves			
New Problem	.013 (.015)	Ok	.060 (.032)
Simplified Problem	.036 (.021)	Positive Feedback	.076 (.042)
Pump	.004 (.007)	Neutral Feedback	.009 (.011)
Hint	.016 (.013)	Negative Feedback	.011 (.010)
Prompt	.020 (.024)	Repetition	.018 (.018)
Forced-Choice Question	.002 (.004)	Attributional Acknowledgement	.006 (.009)
Preview	.002 (.003)	Solidarity Statement	.001 (.002)
Summary	.003 (.005)	Humor	.008 (.015)
Paraphrase	.003 (.004)	General Motivational Statement	.004 (.008)
Example	.006 (.011)	Off-Topic	.061 (.054)
Counter Example	.000 (.001)		
Provide Correct Answer	.013 (.013)		
Direct Instruction	.183 (.063)		
Comprehension Gauging Question	.032 (.023)		

Overall, tutors provided direct instruction, positive feedback, off-topic conversation, and generally broke down the current problem into smaller, more manageable sub-problems. This finding supports the hypothesis consistent with Vygotskian theory (1978). Tutors were found to directly guide learners through problem solving and provide more information to aid in problem solving. In addition, the occurrence of off-topic conversation may support the more nurturing, motivational hypothesis consistent with the INSPIRE Model (Lepper & Woolverton, 2002). Off-topic conversation has previously been found to include rapport building dialogue as well as guidance on more general study skills (Cade, Lehman, & Olney, 2010; Lehman, Cade, & Olney, 2010). This off-topic conversation could be indicative of providing learners with a short mental break when confusion is particularly strong.

Confusion compared to overall dialogue move occurrences. Next, dialogue move occurrences after instances of confusion were compared to dialogue move occurrences overall in expert human tutoring sessions, after instances of other affective states (anxiety, frustration, happiness), after learner questions (common ground, knowledge deficit), and learner answers (correct, partially-correct, vague, error-ridden, and none), with paired-sample *t*-tests. When compared to overall occurrence, there were only differences for repetition ($t(48) = 2.13$, $p = .039$), hint ($t(48) = 2.44$, $p = .018$), new problem ($t(48) = 1.66$, $p = .104$), preview ($t(48) = 3.23$, $p = .002$), counter example ($t(48) = 2.31$, $p = .025$), and provide correct answer ($t(48) = 1.64$, $p = .107$). Repetition ($M = .016$, $SD = .016$), hint ($M = .013$, $SD = .009$), and provide correct answer ($M = .011$, $SD = .008$) all occurred more frequently after confusion than overall, whereas new problem ($M = .015$, $SD = .013$), preview ($M = .003$, $SD = .004$), and counter example ($M = .000$, $SD = .001$) all occurred less frequently after confusion (see Table 2 for after confusion descriptives).

The comparison of dialogue moves after confusion to the overall occurrence of dialogue moves shows that tutors generally continue with typical tutorial instruction after instances of learner confusion. This finding supports systems like ITSpoke (Forbes-Riley & Litman, 2011) that handle confusion similarly to incorrect answers. However, it is important to note some of the differences. The increased occurrence of hints and providing the correct answer provides some support for the approach proposed by VanLehn et al. (2003). In the VanLehn et al. (2003) approach the tutor only provides

the correct answer after the learner has made multiple attempts to solve the problem on their own with only minimal guidance (e.g., hints) from the tutor. In addition, the less frequent occurrence of confusion is notable. This suggests that tutors do not abandon a problem because it is challenging for the learner, but instead tutors may allow learners to remain in a state of confusion and work towards resolving their confusion.

Confusion compared to other affective states. Dialogue move occurrences (i.e., proportion of occurrence) after confusion were next compared to occurrences after the three other most frequently occurring affective states. This analysis was conducted to determine if tutors handled all emotions, particularly negatively-valenced emotions, in a similar manner or if different strategies were adopted. Prompts were found to be the only dialogue move occurrence that differed between confusion and anxiety, with confusion having more prompts than anxiety ($t(47) = 1.81$, $p = .077$, $M = .017$, $SD = .022$). This suggests that aside from follow-up questions with fairly simple desired responses (e.g., one or two words or a key phrase), tutors handled anxiety and confusion in a similar manner. This could be due to the fact that learners' uncertainty about how to proceed could trigger anxiety in this real world context. In other words, being uncertain about how to proceed with the current math problem, for example, could have negative real world consequences such as failing a test or failing a course. Thus, in the present context, instances of anxiety could be highly related to the same events that trigger confusion.

Prompts were also found to occur more after confusion than frustration ($t(26) = 3.30$, $p = .003$, $M = .013$, $SD = .018$) and happiness ($t(46) = 1.79$, $p = .080$, $M = .016$, $SD = .022$). Neutral feedback ($t(26) = 2.20$, $p = .037$, $M = .004$, $SD = .008$) and counter example ($t(26) = 1.79$, $p = .086$, $M = .000$, $SD = .000$) were also found to occur more

after confusion than frustration, whereas solidarity statement ($t(26) = 2.10$, $p = .045$, $M = .004$, $SD = .008$) and general motivational statement ($t(26) = 2.02$, $p = .054$, $M = .010$, $SD = .021$) occurred more after frustration than confusion. Tutors appeared to deploy more motivational and supportive statements after instances of frustration than confusion. It may be the case then that the nurturing component of the INSPIRE model (Lepper & Woolverton, 2002), may be more applicable to instances of frustration than confusion. Tutors may find it more productive to provide tutorial instruction in line with the strategies suggested by VanLehn et al. (2003) and Vygotsky (1978) after instances of confusion.

Humor occurred more following happiness than confusion ($t(46) = 3.01$, $p = .004$, $M = .013$, $SD = .019$) as well as tutor off-topic conversation ($t(46) = 3.34$, $p = .002$, $M = .096$, $SD = .076$) and preview ($t(46) = 2.48$, $p = .017$, $M = .003$, $SD = .007$). On the other hand, positive feedback ($t(46) = 1.75$, $p = .086$, $M = .065$, $SD = .037$), repetition ($t(46) = 1.97$, $p = .055$, $M = .015$, $SD = .017$), forced-choice questions ($t(46) = 2.77$, $p = .008$, $M = .001$, $SD = .003$), simplified problem ($t(46) = 1.78$, $p = .081$, $M = .031$, $SD = .024$), provide correct answer ($t(46) = 2.80$, $p = .007$, $M = .008$, $SD = .008$), and direct instruction ($t(46) = 2.10$, $p = .042$, $M = .159$, $SD = .063$) all occurred more after confusion than happiness. Confusion and happiness seem to contrast each other in terms of tutor dialogue moves. Tutors continue with tutorial instruction on the current problem after confusion, whereas they view happiness as an opportunity to build rapport (humor, off-topic conversation), discuss larger learning strategies (off-topic conversation, Cade et al., 2010; Lehman, Cade, et al., 2010), and move on to a new topic (preview).

Confusion compared to learner questions. Confusion can be indicative of uncertainty or a general lack of understanding about how to proceed, which could be a similar state to when learners ask questions. The next set of analyses investigated differences in how tutors responded to learner questions (common ground, knowledge deficit) compared to instances of confusion with a series of paired-sample *t*-tests. Both types of questions were found to differ in terms of tutor dialogue moves when compared to confusion. First, the comparison to common ground questions is considered. Repetition ($t(47) = 1.85$, $p = .071$, $M = .016$, $SD = .017$), pump ($t(47) = 1.92$, $p = .061$, $M = .003$, $SD = .004$), and off-topic conversation ($t(47) = 2.06$, $p = .045$, $M = .044$, $SD = .030$) were all found to occur more frequently after confusion than common ground questions. In contrast, preview ($t(47) = 1.75$, $p = .088$, $M = .003$, $SD = .007$), counter example ($t(47) = 2.19$, $p = .034$, $M = .001$, $SD = .001$), and positive feedback ($t(47) = 2.25$, $p = .029$, $M = .083$, $SD = .042$) all occurred more frequently after common ground questions than confusion.

Next, the comparison to knowledge deficit questions is explored. There were three similarities between the comparison of confusion to common ground questions and the comparison to knowledge deficit questions. Repetition ($t(44) = 3.53$, $p = .001$, $M = .013$, $SD = .016$) and pumps ($t(44) = 3.04$, $p = .004$, $M = .003$, $SD = .004$) occurred more after confusion than knowledge deficit questions, whereas positive feedback ($t(44) = 2.71$, $p = .010$, $M = .062$, $SD = .035$) occurred more after knowledge deficit questions than confusion. However, knowledge deficit questions also differed from common ground questions when compared to confusion. Specifically, knowledge deficit questions were less likely to be followed by hints ($t(44) = 1.99$, $p = .053$, $M = .011$, $SD = .013$) and

providing the correct answer ($t(44) = 1.71$, $p = .095$, $M = .009$, $SD = .011$) than confusion. The main difference between how tutors handled learner confusion and learner questions, then, was that tutors continued to ask question and progress with the tutoring session more after confusion than when a learner asked a question.

Confusion compared to learner answers. Finally, dialogue move occurrences after confusion were compared to dialogue move occurrences after learner answers (correct, partially-correct, vague, error-ridden, none). Learner correct answers and different types of incorrect answers (partially-correct, vague, error-ridden, none) were all considered in the present analyses. This analysis was performed in light of the UNC-ITSpoke uncertainty intervention that treats instances of uncertainty as similar to incorrect responses (Forbes-Riley & Litman, 2011). The next set of analyses investigated whether dialogue move occurrences after learner answers (correct, partially-correct, vague, error-ridden, none) differed from occurrences after confusion with paired-sample *t*-tests. There were significant differences between all types of answers, with one exception, which suggests that the approach to treat confusion and uncertainty as equivalent to incorrect answers may not be entirely appropriate. Interestingly, there were no significant differences in the way that tutors handled vague answers and the way in which they handled confusion (p's $> .1$). It may be the case then that confusion is only similar to certain types of incorrect answers. Next, the significant differences between confusion and each answer type are considered.

Overall, confusion was not handled in a similar manner to correct answers. This finding is expected. Confusion paired with a correct answer may be evidence that the learner does not fully understand why it is the correct answer or may have even merely

guessed the correct answer. Tutor dialogue moves after confusion were more typical of what would be expected for an incorrect answer. For example, negative feedback more frequently followed confusion ($t(48) = 2.05$, $p = .046$, $M = .008$, $SD = .006$), whereas positive feedback more frequently followed correct answers ($t(48) = 3.97$, $p < .001$, $M = .089$, $SD = .046$). In addition, tutors more frequently used hints ($t(48) = 1.77$, $p = .083$, $M = .013$, $SD = .011$) and provided the correct answer ($t(48) = 1.66$, $p = .104$, $M = .011$, $SD = .008$) after confusion to seemingly work towards resolving misconceptions and gaps in knowledge. In contrast, repetition of the learner's answer ($t(48) = 3.57$, $p = .001$, $M = .022$, $SD = .019$) and counter examples ($t(48) = 1.84$, $p = .071$, $M = .001$, $SD = .001$) were used more after a correct answer than confusion, possibly in an effort to reinforce learners' accurate knowledge. It is interesting to note that off-topic conversation occurred more frequently after confusion than correct answers ($t(48) = 2.35$, $p = .023$, $M = .046$, $SD = .029$). This may indicate that tutors sometimes handle confusion by allowing learners to take a quick mental break or to discuss larger learner strategies.

Both partially-correct answers and confusion were handled by tutors in ways that imply that the learner has inaccurate knowledge; however, the two instances were still handled differently. Overall, the difference involved follow-up questions to a partially-correct answer as opposed to more general information and motivation after confusion. Specifically, prompts ($t(48) = 2.01$, $p = .050$, $M = .026$, $SD = .034$) and pumps ($t(48) = 1.73$, $p = .091$, $M = .006$, $SD = .009$) more frequently followed partially-correct answers than confusion. In contrast, direct instruction ($t(48) = 1.85$, $p = .071$, $M = .168$, $SD = .068$) and general motivational statements ($t(48) = 1.80$, $p = .078$, $M = .003$, $SD = .004$) occurred more frequently after confusion. In addition, counter examples ($t(48) = 1.85$, $p =$

.071, M = .001, SD = .002) and providing the correct answer (t(48) = 2.06, p = .045, M = .016, SD = .012 occurred more frequently after partially-correct answers than confusion. Both of these dialogue moves also differed when confusion was compared to correct answers. However, providing the correct answer was more likely to occur after partially-correct answers, which differed from the pattern found for correct answers. It was once again the case that off-topic conversation was more likely to occur following confusion than a partially-correct answer (t(48) = 1.80, p = .078, M = .050, SD = .038). This finding paired with the finding for general motivational statement may suggest that tutors feel instances of confusion require more support than simply answering incorrectly, at least when the response is partially-correct.

The pattern of findings for the comparison of confusion to error-ridden answers was similar the comparison to partially-correct answers. Specifically, prompts (t(47) = 2.03, p = .048, M = .025, SD = .027), providing the correct answer (t(47) = 1.95, p = .057, M = .018, SD = .017), and counter examples (t(47) = 1.71, p = .092, M = .001, SD = .006) more frequently occurred after error-ridden answers, whereas off-topic conversation occurred more frequently after confusion (t(47) = 3.23, p = .002, M = .036, SD = .028). There were also two findings unique to the error-ridden answer comparison. Presenting a new problem occurred more frequently occurred after confusion (t(47) = 2.20, p = .033, M = .010, SD = .014), whereas negative feedback occurred more frequently after an error-ridden answer than confusion (t(47) = 4.30, p < .001, M = .021, SD = .017). Overall, the pattern of findings in this comparison once again suggests that instances of confusion are not handled by tutors in the same manner as incorrect answers.

Last, tutor dialogue move occurrences following no answers were compared to those following confusion. Interestingly, the pattern of tutor dialogue moves generally contrasted with the pattern found for correct answers. In this comparison, positive feedback was found to occur more frequently after confusion ($t(38) = 1.86$, $p = .070$, $M = .063$, $SD = .043$), while hints ($t(38) = 2.08$, $p = .044$, $M = .025$, $SD = .022$) and providing the correct answer ($t(38) = 3.34$, $p = .002$, $M = .022$, $SD = .018$) occurred more after no answer. In each instance the opposite pattern was found when confusion was compared to correct answers. This pattern supports the notion that confusion is triggered by contradictions between the learners' current knowledge and the current information being presented or anomalous information that does not fit with the general pattern, as proposed by cognitive disequilibrium theory (Festinger, 1957; Graesser et al., 2005; Piaget, 1952). This pattern supports cognitive disequilibrium theory because confusion is not being treated as the absence of knowledge (i.e., no answer), but as inaccurate knowledge. In other words, learners have to know some amount of information, accurate or inaccurate, to be confused. In addition, off-topic conversation ($t(38) = 2.26$, $p = .030$, $M = .045$, $SD = .040$) and attributional acknowledgements ($t(38) = 2.08$, $p = .044$, $M = .003$, $SD = .007$) occurred more frequently after confusion than after no answer. Once again, it appears that more motivational dialogue moves are employed after instances of confusion than incorrect answers. This could suggest that tutors view that learners may need more motivation to persist through instances of confusion than when they simply have inaccurate knowledge.

Overall, it was not the case that instances of confusion were treated the same as when learners gave incorrect answers as in UNC-ITSpoke (Forbes-Riley & Litman,

2011). However, confusion was handled by tutors in the same manner as vague answers. Confusion differed from wrong answers in that tutors provided more motivational statements, presented new problems, and provided the correct answer more when learners were confused. The use of more motivational statements provides support for the nurturing component of the INSPIRE model (Lepper & Woolverton, 2002). When learners answered incorrectly, on the other hand, tutors were more likely to continue breaking down the problem (e.g., simplified problem, hints, prompts, etc.).

Confusion × Cognitive State. Instances of confusion were separated based on the co-occurring cognitive state and differences were investigated between confusion-cognitive state pairs with paired-sample *t*-tests. Confusion paired with answers, questions, and metacomments were considered for the present analyses based on previous research that showed confusion significantly co-occurred with these learner cognitive states (Lehman et al., 2010). Tables 3-6 show the descriptive statistics for each significant comparison of confusion-cognitive state pairs. Pair 1 and Pair 2 in Table 3 refer to each confusion-cognitive state pair in the analysis. The first row of descriptive statistics in Table 3, for example, shows the comparison of Confusion-Knowledge Deficit Question (Pair 1) with Confusion-Error-Ridden Answer (Pair 2). There were not any significant differences for positive feedback, solidarity statement, example, counter example, or comprehension gauging question (*p's* > .1).

The significant findings are organized by the tutorial dialogue moves that differed significantly between confusion-cognitive state pairs. It is important to note the low proportional occurrence of dialogue moves follow each confusion-cognitive state pair (see Tables 3-6). The overall low proportional occurrence is due to the fact that there

were many shared tutor dialogue moves following instances of confusion-cognitive state pairs. In other words, tutors handled the different confusion-cognitive state pairs in an overall similar manner. Thus, the present analysis was a fine-grained examination of the differences in how expert human tutors handled learner confusion.

Tutor feedback dialogue moves. Tutor feedback was considered first (see Table 3 for significant differences). Confusion paired with error-ridden answers was more frequently followed by negative feedback than when paired with knowledge deficit questions ($t(11) = 1.98$, $p = .074$) and less frequently followed by neutral feedback than when paired with both correct ($t(11) = 1.95$, $p = .078$) and partially-correct answers ($t(11) = 2.09$, $p = .061$) as well as when paired with metacomments ($t(12) = 1.98$, $p = .071$). This pattern suggests that the appropriate level of feedback was given to error-ridden answers when confusion was also present. This is consistent with a previous finding that the expert tutors in the present corpus provide discriminating feedback based on the quality of learner responses (D'Mello, Lehman, & Person, 2010).

Confusion paired with metacomments was followed more frequently by negative feedback than when paired with knowledge deficit questions ($t(13) = 2.00$, $p = .067$) and partially-correct answers ($t(11) = 1.93$, $p = .080$). This finding was somewhat confusing given that negative feedback would be unexpected following a metacomment (e.g., *I don't know*, *I got it*). One explanation could be that instances of comprehension (e.g., *I got it*) weren't separated from lack of comprehension (e.g., *I don't know*) in the present corpus. It is the case, however, that learners are generally inaccurate in their assessment of their own comprehension (Dunlosky & Lipko, 2007; Glenberg & Epstein, 1985; Graesser et al., 2009; Hacker et al., 2009). In other words, when learners said "I got it,"

they most likely did not fully understand the current concept. Thus, it could be that tutors presented a new or challenging problem after learners stated that they understand the current concept and then received negative feedback when they were unable to correctly solve the new problem.

Table 3
Mean (SD) of Feedback Dialogue Moves following Instances of Confusion Types

	Feedback			
	Neutral		*Negative*	
	Pair 1	Pair 2	Pair 1	Pair 2
KD vs. ER			.005 (.015)	.021 (.023)
KD vs. MC			.005 (.014)	.020 (.027)
CA vs. ER	.016 (.019)	.003 (.007)		
PC vs. ER	.017 (.023)	.003 (.007)		
PC vs. MC			.005 (.012)	.024 (.027)
ER vs. MC	.001 (.003)	.013 (.022)		

Notes. CG = Common Ground Question, KD = Knowledge Deficit Question, CA = Correct Answer, PC = Partially-Correct Answer, VG = Vague Answer, ER = Error-Ridden Answer, MC = Metacomment.

Tutor motivational dialogue moves. Second, motivational dialogue move comparisons were conducted for each confusion-cognitive state pair (see Table 4 for significant differences). There were three main findings for tutor motivational dialogue moves. First, tutors differed in the way they handled confusion paired with each type of learner question. Tutors more frequently used general motivational statements when confusion was paired with a common ground question ($t(14) = 1.82$, $p = .090$), whereas off-topic conversation was used more after a knowledge deficit question ($t(14) = 1.87$, $p = .083$). This pattern may indicate that tutors want to encourage learners to have more confidence in their knowledge and not feel the need to confirm it with the tutor (i.e., in a common ground question).

Confusion paired with both correct and vague answers also differed from confusion paired with common ground questions. When paired with a correct answer, tutors more frequently followed with repetition compared to pairs with common ground questions ($t(13) = 2.62$, $p = .041$) and vague answers ($t(12) = 2.67$, $p = .020$). Confusion paired with partially-correct answers was also followed more by repetition than pairs with metacomments ($t(11) = 2.10$, $p = .060$). Similar to the finding for common ground question pairs, this may be indicative of tutors attempting to reinforce learners' accurate knowledge.

Table 4
Mean (SD) of Motivational Dialogue Moves following Instances of Confusion Types

	Motivational Statements									
	Attributional Acknowledge		*Humor*		*Repetition*		*General Motivational Statement*		*Off-Topic*	
	Pair 1	Pair 2	Pair 1	Pair 2	Pair 1	Pair 2	Pair 1	Pair 2	Pair 1	Pair 2
CG vs. KD							.004 (.009)	.000 (.000)	.059 (.050)	.033 (.031)
CG vs. CA					.015 (.016)	.039 (.047)				
CG vs. VA	.010 (.018)	.002 (.008)								
CA vs. VA					.047 (.046)	.020 (.022)				
PC vs. MC					.021 (.032)	.007 (.018)				
VA vs. ER	.000 (.000)	.006 (.011)								
VA vs. MC			.006 (.012)	.001 (.004)						

Notes. CG = Common Ground Question, KD = Knowledge Deficit Question, CA = Correct Answer, PC = Partially-Correct Answer, VG = Vague Answer, ER = Error-Ridden Answer, MC = Metacomment.

Last, confusion paired with vague answers was less frequently followed by attributional acknowledgments compared to pairs with error-ridden answers ($t(12) = 1.98$, $p = .072$) and common ground questions ($t(20) = 1.82$, $p = .084$), but more frequently followed by humor than metacomments ($t(14) = 2.02$, $p = .063$). Attributional acknowledgments are a dialogue move that tutors can employ to remove responsibility for confusion, struggles, and failures from learners' knowledge and skills. In other words, answering incorrectly (error-ridden answer) or being unsure about their knowledge (common ground question) is due to the difficulty of the concept, for example, which is consistent with strategies hypothesized by attribution theory (Batson et al., 1995; Heider, 1958; Weiner, 1986). However, it appears that tutors only employ this dialogue move when learners have stated actual knowledge when confused, but not when learners make a vague or incoherent response.

Tutor question dialogue moves. Tutor questions were investigated and three significant patterns were found (see Table 5). First, confusion paired with learner questions was followed by more direct questions than when confusion was paired with incorrect answers. Specifically, common ground questions were followed more by forced-choice questions ($t(20) = 1.94$, $p = .066$) and knowledge deficit questions were followed more by prompts ($t(11) = 2.84$, $p = .016$) than vague answers. In contrast, hints, pumps, simplified problems, and new problems less frequently occurred after learner questions paired with confusion. For common ground question pairs, hints ($t(15) = 2.33$, $p = .034$) and pumps ($t(15) = 1.92$, $p = .074$) occurred more after partially-correct answer pairs, hints also occurred more following error-ridden answer pairs ($t(16) = 1.74$, $p = .101$), and simplified problems occurred more after vague answer pairs ($t(20) = 1.73$, $p =$

.099). For knowledge deficit question pairs, hints ($t(13) = 2.85$, $p = .014$) and new problems ($t(13) = 1.83$, $p = .091$) occurred more after metacomments and new problems were also posed more following error-ridden answer pairs ($t(11) = 2.62$, $p = .024$) and common ground questions ($t(14) = 2.07$, $p = .057$). This pattern suggests that tutors may employ strategies consistent with both the more direct guidance of Vygostskian theory (1978) and the more hands-off guidance strategy proposed by VanLehn et al. (2003); they just deploy these strategies in different contexts.

The second significant pattern was that confusion-metacomment pairs were more frequently followed by prompts (Partially-Correct: $t(11) = 2.71$, $p = .020$) and simplified problems (Error-Ridden: $t(12) = 1.90$, $p = .082$) compared to incorrect answers, but were less frequently followed by posing a new problem (Partially-Correct: $t(11) = 1.87$, $p = .089$). This pattern suggests that tutors stay on task when learners state that they do or do not understand the current topic; however, they break down the current problem into smaller sub-problems that will be more manageable for the learner. The third significant pattern involved comparing learner response types paired with confusion to each other. The significant differences all involved comparisons to vague answer pairs, with vague answer pairs being followed less frequently by prompts and more frequently by simplified problems. Specifically, correct ($t(12) = 2.57$, $p = .024$) and error-ridden answer pairs ($t(12) = 1.92$, $p = .079$) were more frequently followed by prompts, whereas partially-correct ($t(16) = 2.31$, $p = .035$) and error-ridden answer pairs ($t(12) = 3.01$, $p = .011$) were less frequently followed by simplified problems.

The comparison of confusion-vague answer pairs to confusion error-ridden answer pairs was particularly interesting. In both instances the learner has responded

incorrectly and is visibly confused; however, tutors handled the situations differently. It may be the case that when learners provide error-ridden answers the tutor can easily diagnose the specific error or misconception and use a more direct question to correct the error, whereas when learners provide a vague answer, tutors are unsure about the specific error.

Table 5

Mean (SD) of Tutor Question Dialogue Moves following Instances of Confusion Types

		Tutor Questions											
		Forced-Choice		*Hint*		*Prompt*		*Pump*		*New Problem*		*Simplified Problem*	
		P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2
CG vs. KD	M									.01	.01		
	SD									.02	.01		
CG vs. PC	M			.01	.03			.00	.01				
	SD			.02	.03			.01	.02				
CG vs. VG	M	.01	.00									.03	.05
	SD	.02	.00									.03	.04
CG vs. ER	M			.02	.03								
	SD			.02	.02								
KD vs. VG	M					.03	.01						
	SD					.03	.02						
KD vs. ER	M									.01	.03		
	SD									.02	.03		
KD vs. MC	M			.01	.03					.01	.02		
	SD			.01	.03					.01	.03		
CA vs. VG	M					.04	.01						
	SD					.10	.02						
CA vs. MC	M			.01	.03								
	SD			.02	.03								
PC vs. VG	M											.04	.06
	SD											.03	.04
PC vs. MC	M					.01	.02			.02	.01		
	SD					.02	.03			.04	.03		
VG vs. ER	M					.01	.03					.06	.03
	SD					.02	.05					.04	.02
ER vs. MC	M											.03	.04
	SD											.02	.04

Notes. P1 = Pair 1, P2 = Pair 2, CG = Common Ground Question, KD = Knowledge Deficit Question, CA = Correct Answer, PC = Partially-Correct Answer, VG = Vague Answer, ER = Error-Ridden Answer, MC = Metacomment.

Tutor instructional dialogue moves. Finally, tutor instructional dialogue move differences were investigated (see Table 6 for significant differences). There were three main patterns that emerged from these analyses. The first main pattern involved the use of the preview and summary dialogue moves. Tutors more frequently used preview after confusion paired with a vague answer than when paired with a partially-correct answer ($t(16) = 1.74$, $p = .100$), whereas summary was more frequently used after metacomment pairs compared to common ground question pairs ($t(18) = 2.25$, $p = .037$). The second pattern revealed that confusion paired with knowledge deficit questions ($t(13) = 1.81$, $p = .093$) and vague answers ($t(14) = 2.92$, $p = .011$) were more frequently followed by tutors providing the correct answer than metacomment pairs. Both knowledge deficit questions and vague answers represent a gap in learners' knowledge, as opposed to inaccurate knowledge. It seems then that tutors are likely to provide the correct answer to address the knowledge gap, but do not provide the correct answer as a method to correct inaccurate knowledge.

Table 6
Mean (SD) of Tutor Instruction Dialogue Moves following Instances of Confusion Types

	Tutor Instruction							
	Preview		*Summary*		*Provide Correct Answer*		*Direct Instruction*	
	Pair 1	Pair 2	Pair 1	Pair 2	Pair 1	Pair 2	Pair 1	Pair 2
CG vs. KD							.158 (.06)	.204 (.08)
CG vs. PC							.151 (.05)	.126 (.07)
CG vs. MC			.000 (.00)	.004 (.01)				
KD vs. MC					.026 (.03)	.009 (.02)		
PC vs. VA	.000 (.00)	.004 (.01)						
VA vs. MC					.017 (.02)	.005 (.01)		

Notes. CG = Common Ground Question, KD = Knowledge Deficit Question, CA = Correct Answer, PC = Partially-Correct Answer, VG = Vague Answer, ER = Error-Ridden Answer, MC = Metacomment.

The third pattern involved tutors' use of the direct instruction dialogue move. Direct instruction was less likely to follow common ground question pairs compared to knowledge deficit question pairs ($t(14) = 2.29$, $p = .038$), but more likely to follow common ground question pairs when compared to partially-correct answer pairs ($t(15) = 2.02$, $p = .062$). This pattern is interesting in that it can be viewed as three different levels of understanding co-occurring with confusion. Learners can be confused and missing knowledge (knowledge deficit question), confusion and partially accurate in their current knowledge, or confused and unsure about their current knowledge (common ground question). It appears then that as accurate learner knowledge is less present, tutors become more likely to provide direct instruction to address both the learners cognitive and affective state. This progression would be consistent with strategies recommended by both Vygotskian theory (1978) and VanLehn et al. (2003) in that direct instruction is deployed differentially based on the learner's current ability to solve the problem successfully on their own.

4. Study 3: Pedagogical Confusion Interventions

Learning Activity

The central learning activity consisted of critiquing research case studies to determine whether they exhibit sound scientific methodology or have particular methodological flaws. Participants engaged in a trialogue (three-party conversation) with two animated pedagogical agents (tutor and peer student) to evaluate the scientific merits of the case studies. Note that student agent refers to an animated agent; the human learner is referred to as participant or learner for the remainder of the paper. Critical evaluation of case studies involves scientific reasoning skills such as stating hypotheses, identifying dependent and independent variables, isolating potential confounds in designs, and determining if data support predictions (Halpern, 2003; Roth et al., 2006). During the evaluation of a case study, each agent presented its opinion on the scientific merits of the case study and then invited the participant to intervene. For example, in one trialogue the tutor agent asserted that the control group in a study was flawed whereas the student agent disagreed and asserted that the study contains a flaw, but the control group was not flawed. After both agents presented their respective opinions, the tutor agent then asked the participant whether he or she believed that the particular element of the study was flawed (e.g., control group). After the participant gave his or her opinion, the agents presented a second task for the participant to complete. The second task was targeted at helping the participant to learn the material at a deeper level (i.e., understand *why* the particular element of the study was or was not flawed). The specific details of the trialogues are discussed further below. Altogether, participants completed six trialogues with the two agents.

Method

Participants. Participants were 208 undergraduate students from a mid-south university in the US who participated for course credit. There were 149 females and 59 males in the sample. Participants' age ranged from 18 to 56 (M = 21.6, SD = 6.25). Fifty-nine percent of participants were African American, 3% were Asian, 35% were Caucasian, 2% were Hispanic, and 1% were other. Prior coursework in research methods was not required for participation. Eighty-nine percent of participants had not taken a research methods course and 78% had not taken a statistics course.

Confusion Induction Manipulation. Confusion was experimentally induced with a contradictory information manipulation (D'Mello et al., 2014; Lehman et al., 2013). Contradictions were introduced during trialogues that identified flaws in case studies. This manipulation was achieved by having the tutor and student agents stage a disagreement on a concept and eventually invite the participant to intervene. The contradiction is expected to trigger conflict and force the participant to reflect, deliberate, and decide which opinion has more scientific merit. When participants were invited to intervene, they had to decide if they agreed with the tutor agent, the student agent, both agents, or neither of the agents.

There were three contradictory information conditions. In the *True-True* condition, the tutor agent presented a correct opinion and the student agent agreed with the tutor; this was the no-contradiction control. In the *True-False* condition, the tutor agent presented a correct opinion and the student agent disagreed by presenting an incorrect opinion. In contrast, it was the student agent who provided the correct opinion and the tutor agent who disagreed with an incorrect opinion in the *False-True* condition.

It should be noted that all misleading information was corrected and participants were fully debriefed at the end of the experiment.

Confusion Regulation Intervention Manipulation. The pedagogical interventions were designed to help regulate confusion by triggering learners to stop, reflect, and further deliberate over which agent's opinion was correct and *why* that opinion was correct. Interventions were introduced during trialogues that identified flaws in case studies. The interventions occurred after the participant was asked to intervene and decide which agent's opinion has more scientific merit.

There were four intervention conditions. In the Convince Only condition, participants were required to develop a convincing argument that their diagnosis of the case study was correct. The person to be convinced was either one of the agents (True-False, False-True), both agents when the learner disagrees with both agents (True-True), or a hypothetical person when the learner agrees with both agents (True-True). Constructing a convincing argument is expected to cause learners to stop, reflect, and think more deeply about the case study and concept being discussed. In the Read Only condition, participants were presented with an explanatory text to read. The explanatory text may serve as an aid for confusion regulation by providing participants with more information about the concept being discussed.

In the Convince then Read condition, participants were required to first construct a convincing argument as in the Convince Only condition and then were presented with the same explanatory text as in the Read Only condition. The Convince then Read condition may help participants to regulate their confusion by highlighting their confusion and potential gaps in their knowledge in the task of generating a convincing

argument. Participants would then be expected to read the text more deeply in an effort to resolve their confusion and fill in their knowledge gaps that were highlighted during argument construction. In other words, the activity of constructing an argument would cause participants to reach an impasse or be uncertain about how to the concept is applied to the specific situation, both of which would elicit confusion (Brown & VanLehn, 1980; Caroll & Kay, 1988; VanLehn et al., 2003) and would hopefully trigger beneficial cognitive activities that are necessary for participants to return to a state of cognitive equilibrium (Bjork & Linn, 2006; Festinger, 1957; Graesser et al., 2005; Piaget, 1952).

In the Convince while Read condition, participants were provided with the same text as in the Read Only condition, but the text was available as a resource while they constructed their convincing argument. Similar to the Convince then Read condition, the Convince while Read condition may help participants regulate their confusion by both drawing attention to and providing a resource to resolve any confusion and knowledge gaps that emerge during argument construction. However, in the Convince while Read condition participants were able to immediately use the text to resolve any confusion or knowledge gaps that emerged during argument construction. It has been found that most contradictions and anomalies are ignored, dismissed, or erroneously combined with prior misconceptions (Chinn & Brewer, 1993). Thus, even a small time separation between argument construction and reading the explanatory text could allow participants to not address their confusion and gaps in their knowledge.

Design. There was a mixed-design with confusion induction as a within-subjects factor (True-True, True-False, False-True) and confusion regulation intervention as a

between-subjects factor (Convince Only, Read Only, Convince then Read, Convince while Read).

Participants completed two trialogues in each of the three confusion induction conditions with a different research methods concept discussed in each session (6 in all). Figure 3 shows the overall order of events for the experiment as well as the order of events within one trialogue. Participants completed all six trialogues in one of the confusion intervention conditions. The six research methods concepts were construct validity, control groups, experimenter bias, generalizability, random assignment, and replication. Each concept had an associated research case study that was flawed in one significant aspect (e.g., an inappropriate control group). Order of confusion induction conditions and concepts and assignment of concepts to confusion induction conditions was counterbalanced across participants with a Graeco-Latin Square. Confusion intervention condition was randomly assigned to participants.

Trialogues. Each trialogue consisted of three phases that occurred in the following order: (1) induction, (2) intervention, and (3) post-intervention.

Prior to the first trialogue, the tutor and student agents introduced themselves, discussed their roles, discussed the importance of developing research methods knowledge, and described the learning activity. Participants then began the first of six trialogues (see Figure 3). In each trialogue, participants discussed the case study for one of the research methods concepts with the tutor and student agents. The interface that was used for the trialogues is shown in Figure 4. It consisted of (A) the tutor agent, (B) the student agent, (C) a description of the case study, (D) presentation of explanatory text (optional), (E) a text-transcript of the dialogue history, and (F) a text-box for participants

to enter and submit their responses. The tutor and student agents delivered the content of their utterances via synthesized speech, whereas the participant typed his or her responses.

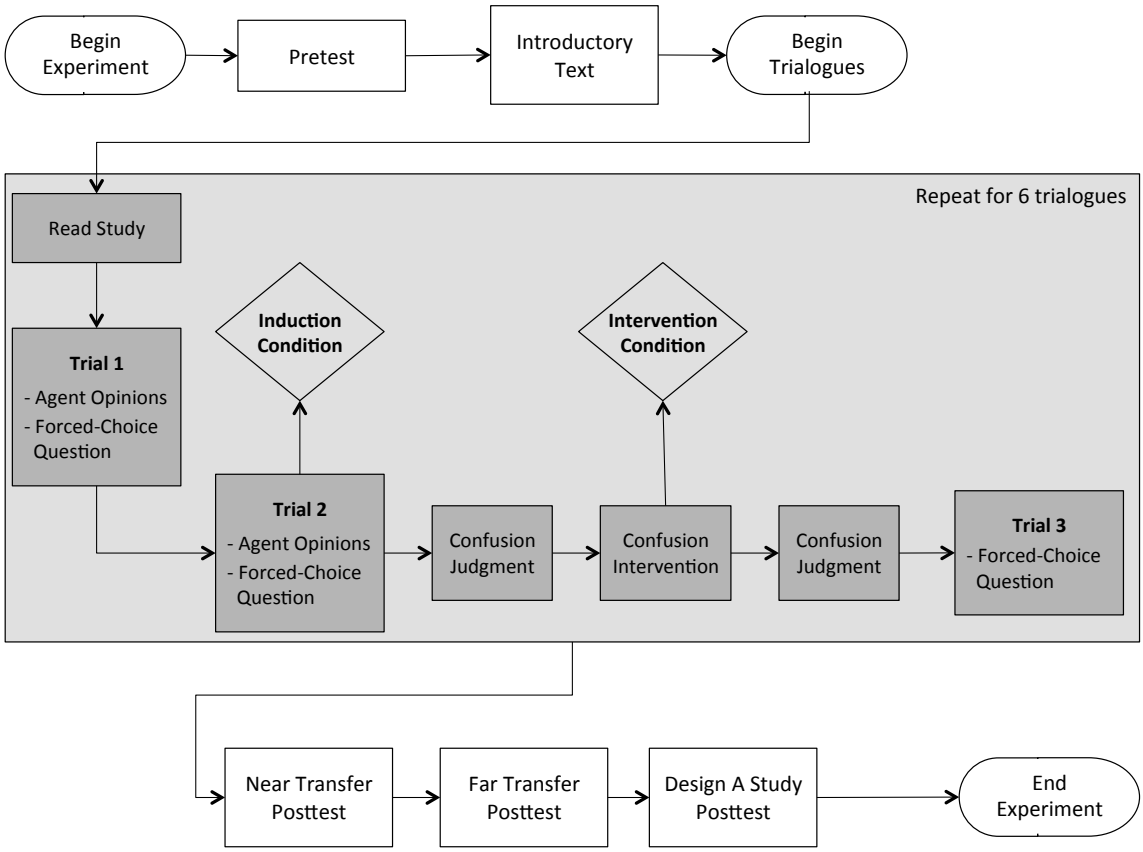

Figure 3. Order of Events for Studies 3 and 4.

Trialogues involved three multi-turn trials, with two trials occurring in the induction phase (before first Confusion Judgment in Figure 3) and one trial occurring in the post-intervention phase (after Confusion Intervention in Figure 3). Each trial involved a tutor posed question to the learner and a learner response, with Trials 1 and 2 also containing each agent stating their opinion about the scientific merits of the case study. For example, in Table 7, turns 7 through 12 represent one-multi-turn trial. The excerpt in

Table 7 is an example trialogue between the tutor agent (Dr. Williams), the student agent (Chris), and the human learner (Bob) from the True-False condition. The agents and Bob are discussing a case study that has an inappropriate control groups as its flaw.

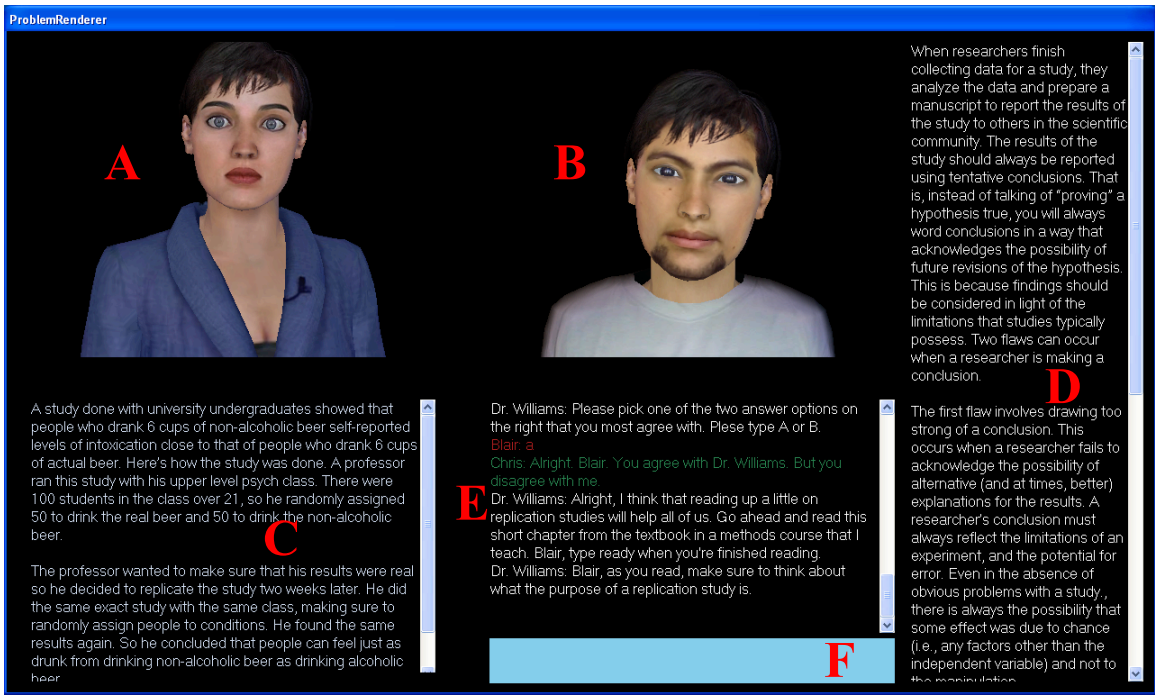

Figure 4. Screenshot of Learning Environment.

The induction phase (turns 1-12) began with a description of the case study that was discussed. The first multi-turn trial (Trial 1) began after the participant read the study. In Trial 1 the trialogue discussed generally whether the study is or is not flawed (turns 1-3). First, the agents asserted their opinion that the study is flawed (turn 1). In Trial 1 the agents agreed and presented correct opinions. Second, the tutor agent asked the participant for his or her opinion via a forced-choice question (turn 2). Third, the participant provided his or her opinion (turn 3). The induction phase then continued with a discussion between the tutor and student agents about a previously identified

misconception when diagnosing the flaw in the case study (turns 4-5) (Lehman et al., 2012).

The second multi-turn trial (Trial 2) in the induction phase began next (turns 7-12). In Trial 2 the trialogue discussed the specific problematic element in the study and why it is problematic. The study in Table 7, for example, has an inappropriate control group. First, the tutor agent provided an opinion and then the student agent either concurred with that opinion or disagreed by providing an alternate opinion (turns 7-8). This is where the confusion induction manipulation occurred in the trialogue (True-True, True-False, False-True). Second, the tutor agent asked the participant for a confusion judgment (turns 9-10). Participants were prompted to indicate whether a classmate would be confused or not confused at this point in the trialogue. The confusion prompt was phrased in this manner because previous research suggests that many learners believe that being in a state of confusion is indicative of poor performance or failure (D'Mello et al., 2014). This perception might reduce participants' tendency to honestly report their level of confusion due to social desirability effects that might be mitigated by asking participants to judge a classmate's level of confusion in the same situation. In addition, previous research has shown that when confusion judgments were phrased in this manner, learners who reported confusion also had longer processing times (Lehman et al., in preparation). Third, the tutor agent asked the participant for his or her opinion via a forced-choice question and the participant provided his or her opinion (turns 11-12).

Table 7

Excerpt of Trialogue of Induction and Post-Intervention Phases from True-False Condition

Turn	Speaker	Dialogue
There was a study where participants took this new diet pill called 'Pounds Off' and reported losing 10 pounds in the first month. None of the participants exercised or ate super healthy or anything, they just acted normally while taking the pill - some people exercised a few times a week, some did not exercise at all, some ate healthy, others did not. The researchers even compared this to a group who didn't take it and just acted how they normally do. This group didn't lose any weight over the course of a month. So Pounds Off works ten times as well as not doing anything. <*Research case study*>		
1	Dr. Williams [tutor agent]	Well Chris and I talked it over and the two of us think that there's something problematic about this study. <*Trial 1 opinions*>
2	Dr. Williams	Bob, do you think there might be a problem with this study? <*Forced-choice question 1*>
3	Bob [human learner]	problem <*Response 1*>
4	Chris [student agent]	The real problem is that diet pills just don't work. Those commercials on television are all made up. So there's no way that this study was done correctly. <*Surface error assert*>
5	Dr. Williams	You have a point, but I think what we need to do is look at the actual methodology of the study to see if there is a problem. <*Surface error correction*>
6	Dr. Williams	Let's talk about the control and the experimental groups they used. <*Advance dialogue*>
7	Dr. Williams	The control and experimental groups were too different. The control group should do something to make the two groups more similar. <*Trial 2 opinion*>
8	Chris	I don't agree with you at all, Dr. Williams. The control group was good. There weren't any differences between the two groups besides the diet pill. <*Trial 2 opinion*>
9	Dr. Williams	Bob, do you think your classmates would be confused? <*Confusion prompt*>
10	Bob	confused <*Confusion judgment*>
11	Dr. Williams	Bob, do you think the control group was appropriate? <Forced-choice question 2> A. The control group is appropriate because it represents the average person in the real world and what he or she would do if they weren't taking the diet pill (i.e., acting normally). <*Incorrect response*> B. The control group would be more appropriate if it were changed so that participants in the control group also took a pill, a pill that wasn't diet related. <*Correct response*>
12	Bob	A <*Response 2*>
		CONFUSION INTERVENTION
13	Dr. Williams	Bob, do you think one of your classmates would be confused? <*Confusion prompt*>
14	Bob	not confused <*Confusion judgment*>
15	Dr. Williams	Alright, Bob, let's consider control groups in another case. <*Introduce new context*>
If you wanted to design a study to test the effectiveness of a computer math tutor, which of the following would be the most appropriate comparison groups. <*Forced-choice question 3*> A. One group use the computer math tutor, a second group use a computer biology tutor, a third group do nothing <*Correct response*> B. One group use the computer math tutor and a second group do nothing <*Incorrect response*> C. One group use the computer math tutor, a second group read the math textbook, and a third group do nothing <*Incorrect response*> D. None of the above <*Incorrect response*>		
16	Bob	A <*Response 3*>
17	Dr. Williams	We've talked about this study for a while now. But we're not going to finish this discussion yet. Ok, let's move on to the next one. <*Closing*>

The intervention phase then occurred after the induction phase was completed. Participants were presented with one of the four confusion interventions in each trialogue (Convince Only, Read Only, Convince then Read, Convince while Read). The

interventions were pedagogical in the sense that they were designed to cause participants to spend additional time reflecting and deliberating about the scientific merits of the case study being discussed. Table 8 shows an example excerpt of the confusion intervention conditions from the *True-False* condition. The human participant (Bob) agreed with the student agent (Chris) and disagreed with the tutor agent (Dr. Williams) in Trial 2 from the induction phase. The trialogue excerpts in Table 8 discuss the same study with an inappropriate control group as in Table 7.

For the Convince Only and Convince then Read conditions, the first task that was posed to participants was to construct a convincing argument (turn 1 in Convince Only and Convince then Read). In the True-False and False-True conditions, the construct argument prompt was delivered by the agent with which the participant disagreed. In the excerpts in Table 7, for example, the construct argument prompt is delivered by the tutor agent (Dr. Williams) because the participant (Bob) response in Trial 2 disagreed with the tutor agent. In the True-True condition, however, the participant either agreed with both or neither of the agents. In either case, the construct argument prompt was delivered by the tutor agent. If the participant disagreed with both agents, then the participant was required to convince both agents that his or her opinion was correct. In the other case, when the participant agreed with both agents, the participant was required to convince a hypothetical new person who disagreed with the participant. The construct argument prompt was identical in the Convince while Read condition with an additional statement to read the explanatory text while constructing the argument (turn 1 in Convince while Read).

Table 8
Excerpt of Trialogue of Intervention Phase from True-False Condition

Turn	Speaker	Dialogue
		CONVINCE ONLY
1	Dr. Williams	I disagree with you, Bob. You need to prove to me that you're right about this study. So put together a convincing argument to get me on your side. <*Construct argument prompt*>
2	Dr. Williams	Bob, remember to think about how participants in the control and experimental groups behaved during the study. <*Hint*>
3	Bob	The control group was appropriate because… <*Convincing argument*>
		READ ONLY
1	Dr. Williams	I disagree with you, Bob. Let's all read over this chapter. <*Introduce explanatory text*>
2	Dr. Williams	Bob, remember to think about how participants in the control and experimental groups behaved during the study. <*Hint*>
To test this hypothesis, you need one or more comparison groups that are not exposed to the treatment… <*Explanatory text*>		
3	Bob	Done <*Finished reading*>
		CONVINCE THEN READ
1	Dr. Williams	I disagree with you, Bob. You need to prove to me that you're right about this study. So put together a convincing argument to get me on your side. <*Construct argument prompt*>
2	Dr. Williams	Bob, remember to think about how participants in the control and experimental groups behaved during the study. <*Hint*>
3	Bob	The control group is better because… <*Convincing argument*>
4	Dr. Williams	Let's all read over this chapter. <Introduce explanatory text>
To test this hypothesis, you need one or more comparison groups that are not exposed to the treatment… <*Explanatory text*>		
5	Bob	Done <*Finished reading*>
		CONVINCE WHILE READ
1	Dr. Williams	I disagree with you, Bob. You need to prove to me that you're right about this study. So put together a convincing argument to get me on your side. Take a look at this chapter while you put together your argument. <*Construct argument prompt + Introduce explanatory text*>
2	Dr. Williams	Bob, remember to think about how participants in the control and experimental groups behaved during the study. <*Hint*>
To test this hypothesis, you need one or more comparison groups that are not exposed to the treatment… <*Explanatory text*>		
3	Bob	The control group was appropriate because… <*Convincing argument*>

The first task for the Read Only condition and the second task for the Convince then Read condition were to read the explanatory text (turn 1 in Read Only and turn 4 in Convince then Read). The texts contained an average of 364 words (*SD* = 41.7) and were adapted from the electronic textbook that accompanies the *OperationARA!* intelligent tutoring system (Halpern et al., 2012). The explanatory text provides participants with more information about the concept being discussed; however, the text did not directly address the case study being evaluated. The tutor agent also presented a hint to

participants in all four confusion intervention conditions (turn 2 in all conditions). The hint was presented prior to the main intervention task (i.e., text read and/or argument construction). The hint was provided to orient participants to the important element of the study (e.g., the experimental and control groups).

Next, all participants began the post-intervention phase (Table 7 turns 13-17). The post-intervention phase served as an assessment of whether confusion had been resolved by the intervention activities for each condition. Participants were first prompted to make a confusion judgment, which was phrased in the same manner as the confusion judgment in Trial 2 (turns 13-14). Next, a forced-choice question was presented that required participants to apply their knowledge of the concept to a new situation (turn 15-16).

Finally, the tutor agent wrapped up the current trialogue with a closing statement and moved on to the next trialogue (turn 17). Misleading information that was delivered in the form of contradictory information was not corrected at the end of each trialogue. Instead, misleading information from all trialogues was corrected by the agents after the posttests had been completed.

Knowledge Tests. Research methods knowledge was assessed with flaw-identification and design-a-study tasks. The flaw-identification task consisted of a description of a previously unseen research study and participants were asked to identify flaw(s) in the study by selecting as many items as they wanted from a list of eight research methods concepts. The list included six concepts that could potentially be flawed (i.e., discussed in the trialogues) and two distractor concepts (i.e., not discussed in the trialogues). Participants also had the option of selecting that there was no flaw in the research study, although each study contained at least one flaw. The pretest involved

identifying the flaws in six case studies in which that each contained one flaw that is discussed in the trialogues (see Appendix N).

The posttest flaw-identification task involved near and far transfer studies. The near transfer studies differed from the studies discussed in the trialogues on surface features only (see Appendix O). Each near transfer study contained one flaw. Each concept discussed during the trialogues had one near transfer study, resulting in six near transfer studies. The far transfer studies differed from the studies discussed in the trialogues on both surface and structural features (see Appendix P). For example, a surface feature difference could be taking a diet pill (original study) versus an acne pill (near transfer study), whereas structural feature differences could be experimental and do-nothing control groups (original study) versus three or more comparison groups all receiving some type of treatment (far transfer study). Each far transfer study contained two flaws, resulting in three far transfer studies in all. Both the near and far transfer studies were presented after all of the trialogues were completed.

The design-a-study task required participants to design a hypothetical research study to test a claim (i.e., "Teachers always tell their students that it is better to study a little but of the course material each day, rather than try to cram all of the studying into the night before the exam.") (see Appendix Q). Participants answered four-alternative, forced-choice (4AFC) multiple-choice questions pertaining to each concept discussed in the trialogues. Each question required participants to make a decision in order to avoid a potential flaw in the study to test the claim about study habits. For example, participants needed to decide how to avoid potential experimenter bias in data collection as well as

the amount and nature of the comparison groups in the study. The design-a-study task was only presented after all trialogues have been completed.

Procedure. Participants were individually tested over a two-hour session. The order of events is shown in Figure 3. First, participants signed an informed consent and data release agreement (see Appendices J and K) and then completed the pretest. Next, participants read a short introductory text on research methods. The introductory text provided participants with a broad overview of the research methods terminology that was discussed during the trialogues.

Participants then began the first of six trialogues. In each trialogue, participants discussed the case study for one research methods concept with the tutor and student agents. Three streams of information were recorded as participants completed the trialogues. First, a video of the participant's face was captured using a webcam that was integrated into the computer monitor. The webcam also recorded all audio generated during the interaction. Second, a video of the participant's screen was recorded using a commercially available screen capture program called Camtasia Studio™. Third, a variety of interaction parameters were automatically recorded in log files. These parameters included the participant's responses (typed responses and response times) and the current state of the interaction (e.g., pretest vs. trialogue). After completing all six trialogues, participants completed the flaw-identification and design-a-study posttests. Participants then interacted with the agents again to have all misleading information corrected. Participants were fully debriefed at the end of the experiment (see Appendix L).

Results and Discussion

There were four sets of dependent measures in the analyses: confusion induction (induction phase), confusion regulation process (intervention phase), confusion regulation outcome (post-intervention phase), and learning outcome measures. A mixed-effects modeling approach was adopted for all analyses due to the repeated measurements and nested structure of the 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 computations.

Linear or logistic models were constructed on the basis of whether the dependent variable was continuous or binary, respectively. The random effects in all analyses were participant and concept. In addition, all models included order as a fixed effect (order of concept presentation) and time on task (for each trialogue) as a fixed effect. Time on task was included because the confusion intervention conditions were not equivalent in time on task. The random effects and fixed effects of order and time on task were consistent across all models (control). Induction condition, intervention condition, and/or induction × intervention were the categorical fixed effects. The comparisons reported for induction condition differences focused on the a priori comparison of each experimental condition to the no contradiction control, so the True-True condition was set as the reference group in all of the models. The comparisons reported for intervention condition differences, on the other hand, compared each intervention condition to all other intervention conditions. One-tailed tests were used for significance testing when the hypothesis specified the

direction of the effect. However, two-tailed tests were used when no a priori predications were made. All significance testing was conducted with an alpha level of .05. The unit of analysis was the case study (or individual trialogue) for all analyses. There were 1234 observations in the present analyses.

Confusion Induction. There were three dependent measures from the induction phase of the trialogue that were used in the present analyses: confusion judgment, Trial 1 response quality, and Trial 2 response quality. Table 9 shows the coefficients for the models along with the mean proportional occurrence of each dependent measure. Mixed-effect logistic regression models were constructed for each dependent measure. For confusion induction, it was hypothesized that when participants were in the *True-False* and *False-True* conditions, they would report more confusion than when in the *True-True* condition based on cognitive disequilibrium theory (Festinger, 1957; Graesser et al., 2005; Piaget, 1952). In other words, the presentation of contradictory information was expected to induce greater levels of confusion (on average) than agreement. The model was significant[1] and supported the hypothesis with participants reporting more confusion when in both experimental conditions than the no-contradiction control condition ($\chi^2(2) = 17.4$, $p < .001$).

Next, response quality on Trials 1 and 2 were investigated. For Trial 1, a significant difference between the *True-True* condition and the experimental conditions (True-False, False-True) was not expected because the agents agreed and provided a correct opinion. Participants were expected to respond similarly when both agents agreed,

[1] Significance of logistic mixed-effects models is generally evaluated by comparing the mixed-model (fixed + random effects) to a random model (random effects only) with a likelihood ratio test. In the present study, significance was evaluated by comparing the full model (fixed effects + control model) to the control model only with a likelihood ratio test.

which in this case is correct, based on previous research (D'Mello et al., 2014; Lehman et al., 2013). The hypothesis for Trial 1 was also supported with a non-significant model ($\chi^2(2) = 3.15$, $p = .207$) showing that the experimental conditions did not differ from the no-contradiction control condition.

However, it was hypothesized that there would be a difference in response quality for Trial 2. Participants were expected to perform less well (on average) when in the *True-False* and *False-True* conditions than when in the *True-True* condition. The assumption was that if participants were confused by the presentation of contradictory information, they would be uncertain about which opinion was correct and would respond incorrectly more frequently (on average) than when the agents agreed with each other. The hypothesis for Trial 2 was also supported. The model was significant with participants responding less accurately in the experimental conditions than in the no contradiction control condition ($\chi^2(2) = 9.18$, $p = .010$). Overall the findings for the confusion induction phase supported the hypothesis for each dependent variable and showed that the presentation of contradictory information by animated pedagogical agents can successfully induce confusion in the minds of learners.

Table 9
Proportional Occurrence of Induction Phase Dependent Measures

	Induction Condition			**Coefficient (B)**	
	Tr-Tr	*Tr-Fl*	*Fl-Tr*	*Tr-Fl*	*Fl-Tr*
Confusion Judgment	.610	.700	.700	**.578**	**.683**
Trial 1	.710	.750	.750	.258	.255
Trial 2	.750	.660	.680	**-.471**	**-.360**

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

Confusion Regulation Process. The confusion regulation process analyses involved dependent measures that occurred during the confusion intervention manipulations. The dependent measures included explanatory text read time (seconds), argument construction time (seconds), argument construction + text read time (seconds), argument length (words), and argument quality (discussed below).

Argument quality assessment. Argument quality was assessed in two ways. First, participant arguments were compared to prototypical correct responses that were created by a content expert. Prototypical correct responses were unique to each of the six research methods concepts discussed during the trialogues. Participant arguments were compared to prototypical correct responses using an inverse word frequency weighted overlap (IWFWO) algorithm. The IWFWO algorithm is a word-matching algorithm in which each overlapped word is weighted on a scale from 0 to 1, relative to its inverse frequency in the English language using the CELEX corpus (Baayen, Piepenbrock, & Gulikers, 1995). The inverse frequency allows for higher weighting of lower frequency, more contextually relevant words (e.g., replication, bias), while higher frequency words (e.g., and, but) are given a lower weighting. Comparisons resulted in a match score between 0 and 1 (1 = perfect similarity). This match score served as the semantic match score dependent variable in subsequent analyses.

The second assessment of participant arguments was through coding by trained judges. There are multiple methods to code scientific arguments (see Sampson & Clark, 2008); however, the most appropriate method for the present study was the scheme presented by Zohar and Nemet (2002), which focuses more on the content quality than the structure (see Appendix R). This coding scheme involved identifying instances of

claims and evidence in the argument. The scheme has been adapted to include the rating of claim and evidence quality (accurate, inaccurate), based on the suggestion of Sampson and Clark (2008). Two human raters coded the arguments for the presence of claims, evidence, the amount of evidence, and the quality of claims and evidence. A subset of the corpus was first coded to compute reliability (*kappa* = .842). The corpus was then divided evenly between the raters for coding. This coding scheme was then used for five argument quality dependent measures: claim quality, evidence quality, amount of evidence, overall presence score, and overall quality score. Claim quality, evidence quality, and amount of evidence were taken directly from the coding scheme. The overall presence score was derived from the coding scheme with scores of 0 (neither claim nor evidence was present, N = 70), 1 (claim or evidence was present, N = 352), and 2 (claim and evidence were present, N = 496). The overall presence score was dummy coded for analyses, such that individual scores were predicted (i.e., 0, 1, or 2). The overall quality score combined the overall presence score with information about the quality of the claim and evidence that were present. This scored divided arguments into Low (N = 599) and High (N = 377). Arguments score as Low were missing the claim, evidence, or both and had an incorrect claim and/or evidence. On the other hand, arguments scored as High included both claims and evidence with at least one being accurate.

Confusion regulation process analyses. There were nine dependent measures for the confusion regulation process: explanatory text read time and argument response time, number of words, semantic match score, claim quality, evidence quality, amount of evidence, overall presence score (0, 1, 2), and overall quality score. Table 10 shows the proportional occurrence for each regulation process dependent measure at the induction

condition and intervention condition levels. Note that standard deviations are not included in Table 10 because some measures are binary and mixed-effects models involve analyses that include nested variables and adjustments based on the random effects. So there is not one standard deviation in the present analyses.

Table 10
Proportional Occurrence of Confusion Regulation Process Dependent Measures

	Induction and Intervention Condition						
	Tr-Tr	*Tr-Fl*	*Fl-Tr*	*Convince Only*	*Read Only*	*CtR*	*CwR*
Convince RT	107	100	103	101		105	
Text RT	97.0	102	98.1		101	97.6	
Text + Convince RT	193	201	185				193
Convince Word Num	37.6	39.5	38.9	36.6		36.6	42.8
Claim Quality	.610	.540	.440	.540		.520	.520
Evidence							
Amount	1.28	1.23	1.22	1.22		1.28	1.22
Quality	.560	.450	.400	.460		.480	.470
Presence Score							
Zero	.090	.060	.080	.070		.060	.110
One	.400	.360	.390	.410		.400	.330
Two	.510	.580	.530	.520		.540	.570
Quality Score	.460	.440	.340	.380		.430	.430
Semantic Match Score	.402	.385	.361	.389		.387	.370

Notes. Tr = True, Fl = False, CtR = Convince then Read, CwR = Convince while Read, RT = response or read time

Three models were constructed for each dependent measure: induction main effect, intervention main effect, and induction × intervention interaction. Cognitive disequilibrium (Festinger, 1957; Graesser et al., 2005; Piaget, 1952) and impasse-driven theories of learning (Brown & VanLehn, 1980; VanLehn et al., 2003) would hypothesize that learners would have increased processing time (i.e., text read time, argument construction + text read time) when in the *True-False* and *False-True* conditions compared to the *True-True* condition. This pattern was predicted because when learners

were in the contradictory information conditions they would exert greater effort in order to resolve their current confusion.

The confusion induction processing time hypothesis was not directly supported because there was not a significant induction condition main effect ($p > .1$). However, there was a significant main effect of confusion induction success (i.e., confusion judgment in the Induction Phase). Participants who reported confusion in the induction phase had longer argument response times ($F(1,610) = 3.20$, $p = .074$) and explanatory text reading times ($F(1,620) = 11.0$, $p = .001$) than those who did not report confusion. This finding is consistent with cognitive disequilibrium (Festinger, 1957; Graesser et al., 2005; Piaget, 1952) and impasse-drive theories of learning (Brown & VanLehn, 1980; VanLehn et al., 2003) in that confusion triggered increased processing time, although not in the manner posed by the specific hypothesis. The absence of an induction condition main effect may stem from the source of confusion for participants. Although participants were not presented with contradictions when in the True-True condition, it is still possible for confusion to occur. It may be the case that participants in the True-True condition were confused because of the difficulty of or lack of familiarity with the material being presented.

The confusion intervention condition was also expected to have an impact on the amount of processing time during the confusion regulation process. There were two primary comparisons conducted. First, the text read time was compared between the Read Only and Convince then Read conditions. The Convince then Read condition was expected to have greater reading times than the Read Only condition because the construction of an argument would presumably highlight confusion and knowledge gaps

that can be addressed by processing the text more deeply. This hypothesis was not confirmed. The results showed that the Read Only and Convince then Read conditions did not significantly differ on text read time ($p > .1$). In fact, the Read Only condition had longer reading times than the Convince then Read condition, although it was only a small difference in time (3.4 seconds). Second, argument construction time was compared between the Convince Only and Convince then Read conditions. The Convince Only and Convince then Read conditions were expected to be similar in processing time because these two conditions are initially presented with the same task. This hypothesis was confirmed. The Convince Only and Convince then Read conditions did not significantly differ in the amount of time spent on argument construction. There was not a significant induction × intervention interaction for processing time measures ($p > .1$).

Argument quality dependent measures were subsequently investigated with linear and logistic regression models. There were no specific hypotheses for how induction condition would impact argument quality. However, significant differences were found between the experimental conditions and the no-contradiction control condition. Overall, participants had lower quality arguments when in the False-True condition compared to the True-True condition. Specifically, when in the False-True condition participants were less likely to make a correct claim ($\chi^2(2) = 11.1$, $p = .004$, B = .712), present correct evidence ($F(2,936) = 5.23$, $p = .006$, B = .139), and more likely to have an overall lower score ($\chi^2(2) = 11.5$, $p = .003$, B = .603). When participants were in the True-False condition they were also less likely to present correct evidence (B = .091), but they were more likely to present both a claim and evidence (presence score of 2) compared to the True-True condition ($\chi^2(2) = 3.95$, $p = .069$, B = 367).

It is interesting that the two experimental conditions differed on argument quality. In particular, the False-True condition generally provided arguments of low quality. It appears to be the case that confusion can trigger longer processing times, but does not necessarily lead to higher quality arguments. In previous experiments it has been suggested that the degree of confusion may be higher in the True-False condition compared to the False-True condition (D'Mello et al., 2014; Lehman et al., 2013). This pattern suggests that when participants are in the False-True condition they are spending more time attempting to resolve their confusion. However, their efforts do not appear to lead to successful confusion resolution.

Similar to argument construction time, the Convince Only and Convince then Read conditions were expected to be similar in length and quality. This hypothesis was confirmed. The Convince Only and Convince then Read conditions did not significantly differ on any of the argument quality measures ($p > .1$). However, the Convince while Read condition was expected to differ on both length and argument quality from both the Convince Only and Convince then Read conditions. Specifically, the Convince while Read condition was expected to have longer and higher quality arguments because of the availability of the text to resolve confusion and knowledge gaps. This hypothesis was not confirmed. The Convince while Read condition also did not significantly differ from the Convince Only and Convince then Read conditions on any of the argument quality measures ($p > .1$). When examining the proportional occurrence of argument quality measures it was, however, the case that the Convince while Read had longer arguments on average and was more likely to present both a claim and evidence (presence score of 2), but the Convince while Read condition was also more likely to present neither a claim

nor evidence (presence score of 0) and had a lower semantic match score to the ideal argument. It was not the case then that providing the explanatory text during argument construction led to higher quality arguments. There was not a significant induction × intervention interaction for argument quality measures ($p > .1$).

The previous analyses showed that the experience of confusion during the induction phase was related to increased processing time. It follows then that confusion induction success may moderate the impact of induction and intervention condition on argument quality. Cognitive disequilibrium (Festinger, 1957; Graesser et al., 2005; Piaget, 1952) and impasse-driven theories of learning (Brown & VanLehn, 1980; VanLehn et al., 2003) would also suggest that the amount of effort put into the confusion regulation task would also impact the confusion resolution process, outcome, and ultimately learning. Argument quality was further investigated by incorporating both confusion induction success (i.e., confusion judgment from induction phase) and the amount of effort put into the regulation task. Table 11 presents the proportional occurrence of each argument quality measure for the induction × confusion (confused, not confused) × intervention × regulation effort (high, low) interaction. Regulation effort was defined as the amount of time (seconds) on the regulation task (i.e., text read, argument construct, text read + argument construct).

The analysis proceeded by dividing the 1216 cases into low vs. high regulation effort cases based on a median split of participants' regulation time for each case study. The median split was performed separately for each intervention condition. There were 608 low regulation effort cases and 608 high regulation effort cases. The interaction was significant for claim quality ($\chi^2(35) = 48.3$, $p = .066$), presence score of one ($\chi^2(35) =$

49.0, $p = .059$), semantic match score ($F(4,936) = 2.76$, $p = .027$), and overall score ($\chi^2(35) = 61.5$, $p = .004$). The interactions were then examined by dividing cases based on confusion induction success (confused, not confused), regulation effort (low, high), and either induction condition or intervention condition. Induction condition or intervention condition were then regressed onto each case separately. The findings for the significant overall score interaction are discussed next. The discussion of findings for the remaining significant interactions can be found in Appendix T.

The patterns for overall score revealed the circumstances under which the two experimental conditions had higher or lower overall scores for argument quality compared to the no-contradiction control condition. Participants were less likely to have a high score when in the True-False condition compared to the True-True condition when they were not confused, had low regulation effort, and were in the Convince Only condition ($\chi^2(2) = 6.09$, $p = .048$, B = 2.06). When participants were confused and in the Convince then Read condition, they were less likely to have an overall high score whether they had low ($\chi^2(2) = 6.99$, $p = .030$, B = 1.99) or high regulation effort ($\chi^2(2) = 14.4$, $p < .001$, B = 2.17). These findings could be expected given that participants were less likely to respond correctly in Trial 2 (see Confusion Induction) and had no additional information provided to correct any erroneous claim or allow for accurate evidence to be included in the argument.

However, there was one case in which participants had higher overall scores when in both experimental conditions ($\chi^2(2) = 15.5$, $p < .001$). When participants were not confused, had high regulation effort, and were in the Convince while Read condition, they had a higher overall score when in both the True-False (B = 34.0) and False-True

Table 11

Proportional Occurrence of Argument Quality Dependent Measures

	True-True			True-False			False-True		
	CO	*CtR*	*CwR*	*CO*	*CtR*	*CwR*	*CO*	*CtR*	*CwR*
Claim Quality									
NC-LR	.790	.370	.380	.330	.400	.170	.600	.360	.400
NC-HR	.710	.630	.530	.330	.670	.690	.250	.330	.440
C-LR	.440	.610	.560	.650	.540	.520	.470	.430	.500
C-HR	.700	.710	.770	.570	.620	.520	.570	.450	.470
Presence Score									
Zero									
NC-LR	.130	.140	.050	.000	.050	.230	.000	.170	.070
NC-HR	.000	.000	.100	.000	.000	.000	.000	.000	.000
C-LR	.070	.160	.250	.210	.080	.110	.160	.090	.230
C-HR	.030	.030	.110	.000	.000	.050	.030	.000	.050
One									
NC-LR	.480	.450	.550	.570	.400	.310	.620	.440	.530
NC-HR	.500	.240	.250	.270	.210	.060	.220	.230	.270
C-LR	.550	.420	.360	.290	.530	.390	.560	.580	.390
C-HR	.430	.370	.210	.380	.320	.340	.260	.390	.260
Two									
NC-LR	.390	.410	.400	.430	.550	.460	.380	.390	.400
NC-HR	.500	.760	.650	.730	.790	.940	.780	.770	.730
C-LR	.380	.420	.390	.500	.390	.500	.280	.330	.390
C-HR	.540	.600	.680	.620	.680	.610	.710	.610	.690
Quality Score									
NC-LR	.390	.320	.300	.140	.400	.080	.460	.220	.330
NC-HR	.600	.670	.500	.450	.570	.820	.330	.460	.450
C-LR	.240	.420	.320	.360	.320	.320	.130	.180	.320
C-HR	.540	.540	.640	.490	.680	.470	.410	.390	.490
Semantic Match Score									
NC-LR	.519	.350	.401	.541	.398	.368	.216	.359	.402
NC-HR	.377	.454	.598	.361	.388	.363	.439	.395	.244
C-LR	.414	.361	.371	.355	.275	.393	.334	.403	.207
C-HR	.288	.410	.380	.437	.443	.329	.389	.418	.436

Notes. CO = Convince Only, CtR = Convince then Read, CwR = Convince while Read, NC = not confused, C = confused, LR = low regulation effort, HR = high regulation effort

conditions (B = 18.9). It is interesting that the conditions under which this pattern occurred was when participants were not confused and had high regulation effort. Given that participants were not confused, it could be assumed that they would have less motivation to engage in the confusion regulation task. This finding necessitates further investigation to determine what motivated participants in these circumstances to engage in the regulation task.

Overall, the findings from the argument quality dependent measures may suggest that it is not necessary for the argument to be of good quality to resolve confusion and help learning. It may simply be necessary to engage in the process of creating the argument (e.g., comparing different perspectives). The outcome of confusion regulation is investigated next to further explore this issue.

Confusion Regulation Outcome. Confusion resolution outcome analyses involved dependent measures that occurred during the post-intervention phase of the trialogues. The dependent measures were confusion resolution (discussed below) and Trial 3 response quality. Table 12 shows the proportional occurrence of each confusion resolution outcome and performance on the Trial 3 forced-choice question. Mixed-effects logistic regressions were constructed for each dependent measure for induction main effect, intervention main effect, and induction × intervention interaction. There were no expected patterns for induction main effect or the induction × intervention interaction. However, there were expected patterns for the intervention main effect. Overall, participants in the most cognitively engaging condition (Convince while Read) were expected to resolve their confusion and respond correctly to the Trial 3 forced-choice question more than participants in the other conditions.

To determine whether confusion was or was not resolved, it was necessary to consider confusion both before and after the intervention. Confusion resolution was defined as the change in confusion from Time 1 (T1, induction phase) to Time 2 (T2, post-intervention phase). There were four possible outcomes: none (not confused at T1 or T2, $N = 332$), resolved (confused at T1, not confused at T2, $N = 422$), unresolved (confused at T1 and T2, $N = 401$), and created (not confused at T1, confused at T2, $N = 77$). Separate mixed-effect logistic regression models were constructed for each outcome (dummy coded).

For the induction condition main effect, there were significant models for None ($\chi^2(2) = 11.2$, $p = .004$) and Unresolved ($\chi^2(2) = 6.71$, $p = .035$), but the models for Resolved and Created were not significant (p's $> .1$). It was the case, however, that both the True-False (.370) and False-True conditions (.343) had higher proportional occurrences of resolved confusion than the True-True condition (.315). Both experimental conditions were less likely to have no confusion than the no contradiction control condition (True-False: B = -.482, False-True: B = -.585) and were more likely to have unresolved confusion (True-False: B = .237, False-True: B = .459). These findings suggest that participants were more likely to remain in a state of confusion when in the experimental conditions, which according to cognitive disequilibrium (Festinger, 1957; Graesser et al., 2005; Piaget, 1952) and impasse-driven theories of learning (Brown & VanLehn, 1980; VanLehn et al., 2003) would suggest that they would not successfully learn the material. However, recent research has shown some evidence that complete

Table 12

Proportional Occurrence of Post-Intervention Phase Dependent Measures

	Induction × Intervention Interaction											
	Convince Only			*Read Only*			*Convince then Read*			*Convince while Read*		
	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Tr	Tr-Fl	Fl-Tr
Confusion Resolution												
None	.290	.310	.280	.330	.210	.260	.270	.240	.200	.390	.240	.220
Resolved	.350	.310	.260	.330	.460	.360	.330	.370	.370	.250	.340	.380
Unresolved	.310	.350	.430	.290	.320	.320	.280	.310	.340	.290	.330	.350
Created	.050	.030	.040	.050	.020	.060	.120	.070	.090	.070	.090	.050
Low Regulation Effort												
None	.380	.410	.270	.430	.200	.250	.300	.290	.270	.410	.280	.250
Resolved	.230	.240	.270	.290	.450	.400	.360	.430	.330	.220	.340	.270
Unresolved	.330	.330	.440	.200	.320	.340	.230	.220	.310	.300	.300	.400
Created	.060	.020	.020	.080	.020	.020	.110	.050	.080	.070	.090	.080
High Regulation Effort												
None	.180	.190	.290	.220	.210	.280	.250	.190	.140	.350	.220	.200
Resolved	.470	.370	.250	.380	.460	.320	.300	.310	.400	.290	.330	.480
Unresolved	.310	.400	.400	.380	.300	.300	.320	.400	.370	.290	.360	.300
Created	.040	.040	.060	.020	.020	.110	.130	.100	.090	.060	.090	.020
Trial 3	.190	.170	.110	.230	.140	.190	.220	.110	.150	.130	.180	.120

Notes. Tr = True, Fl = False

confusion resolution was not necessary for learning to occur (D'Mello & Graesser, in press).

For the intervention main effect, all models were not significant (*p's* > .1) except for the Created model ($\chi^2(3) = 6.96$, $p = .073$). However, when the proportional occurrence of resolved confusion was examined, the following overall pattern was found: Convince Only (.307) < Convince while Read (.323) < Convince then Read (.357) < Read Only (.383). Although not significant, it is still interesting to note that resolved confusion occurred the most in the Read Only condition. Participants in the Convince then Read condition were more likely to have created confusion than those in the Convince Only (B = .898) and Read Only conditions (B = .990). Participants in the Convince while Read condition were also more likely to have created confusion than those in the Read Only condition (B = .620). This finding reveals that confusion was created by the combination of the two regulation tasks. This suggests that neither constructing an argument nor reading the explanatory text alone created confusion, but the combination of the two created confusion for participants. However, it was not the case that these two conditions had more unresolved confusion. Thus, it may have been that some of the participants who reported no confusion in the induction phase were not accurate in their judgment and subsequently became confused when they were asked to apply their knowledge and presented with additional information about the concept.

The induction × intervention interaction was only significant for the None model ($\chi^2(11) = 19.0$, $p = .060$). To investigate the interaction, models investigating the induction main effect were conducted separately for each intervention condition. This revealed a significant model only for the Convince while Read condition ($\chi^2(2) = 10.4$, *p*

= .006) with participants in both experimental conditions (True-False: B = -.909, False-True: B = -1.09) being less likely to have no confusion than the no-contradiction control condition. This finding reveals that the overall induction condition main effect was particularly attributable to the Convince while Read condition.

An important component to confusion resolution is whether or not the participant put in the effort needed to resolve their confusion. Regardless of the method of confusion induction or confusion intervention provided, if the participant does not put forth the effort, then confusion resolution is unlikely to occur. To address this issue, the induction × intervention × regulation effort interaction was investigated. Regulation effort was defined in the same manner as in the analyses for the confusion regulation process analyses. The interaction term was only significant for the None ($\chi^2(23) = 38.6$, $p = .022$) and Resolved models ($\chi^2(23) = 36.8$, $p = .034$), but not the Unresolved and Created models (p's > .1).

The interaction was examined by regressing confusion resolution for the low and high regulation effort cases separately for each intervention condition. There were two significant models when no confusion occurred and in both models the no-contradiction control condition had a higher occurrence than the experimental conditions. The two models were the cases when participants had low regulation effort and were in the Read Only condition ($\chi^2(2) = 7.00$, $p = .030$, True-False: B = -1.50, False-True: B = -1.13) and when participants had high regulation effort and were in the Convince while Read condition ($\chi^2(2) = 8.08$, $p = .018$; True-False: B = -1.29, False-True: B = -1.43). These findings have once again narrowed the conditions under which the True-True condition had a higher occurrence of no confusion.

The significant models for the Resolved cases revealed the conditions under which the False-True condition differed from the True-True condition. In both instances participants had high regulation effort. Participants in the Convince Only condition had less resolved confusion when in the False-True condition ($\chi^2(2) = 4.86$, $p = .088$, B = -1.05), but those in the Convince while Read condition had more resolved confusion when in the False-True condition ($\chi^2(2) = 5.68$, $p = .058$, B = 1.05). This finding is consistent with the original hypothesis that participants would have more confusion in the Convince while Read condition. The addition of regulation effort as a moderator causes this finding to be even more consistent with cognitive disequilibrium (Festinger, 1957; Graesser et al., 2005; Piaget, 1952) and impasse-driven theories of learning (Brown & VanLehn, 1980; VanLehn et al., 2003). Participants had to put in effort to benefit from the regulation task and resolve their confusion.

Mixed-effects logistic regression models were also constructed to investigate response quality on Trial 3. There was not a hypothesis for the impact of induction condition; however, the Convince while Read condition was expected to perform better on the Trial 3 forced-choice question. This prediction was based on the same reasoning as the prediction that the Convince while Read condition would have the most resolved confusion. None of the models were significant for Trial 3 response quality (p's > .1). When proportional occurrence of correct responses was investigated, the following pattern was found: Convince while Read (.143) < Convince Only (.157) = Convince then Read (.157) < Read Only (.187). It is interesting to note that once again the Read Only condition seems to have the most successful outcome from the confusion regulation task.

Learning Outcomes. Learning outcome analyses included the three posttest dependent measures: near transfer, far transfer, and design-a-study tasks. Performance on the transfer tasks was assessed by hits (i.e., correctly identify flaw in a study), whereas performance on the design-a-study task was assessed by selecting the correct multiple-choice answer option. Mixed-effects logistic regression models (1 = correct, 0 = incorrect) were constructed for each posttest dependent measure. When a significant model for hits was found on the near or far transfer tasks, mixed-effects linear regression models were constructed to investigate false alarms (i.e., incorrectly identify flaw in a study) to determine if performance was due to guessing. Learning outcome analyses consisted of three phases: main effects (induction, intervention), induction × intervention interaction, and the impact of confusion induction success and regulation effort.

Previous research has revealed that the presentation of contradictions alone has not been sufficient to increase learning (D'Mello et al., 2014; Lehman et al., 2013). Therefore, a significant induction condition main effect was not expected in the present analyses. However, the confusion intervention conditions were expected to differentially impact performance on the posttests. The expected pattern of performance was Convince Only < Read Only < Convince then Read < Convince while Read. Participants in the Convince while Read condition were expected to perform better than those in all other intervention conditions because participants had the opportunity to reflect on the scientific merits of the case study, deliberate over which opinion holds more merit, and use the explanatory text to address confusion and knowledge gaps that emerge during argument construction. In other words, participants in the Convince while Read condition had the greatest opportunity to successfully resolve their confusion and thus learn the

material more deeply (D'Mello & Graesser, in press; D'Mello et al., 2014; VanLehn et al., 2003). The remainder of the pattern represents a reduction in the potential to successfully resolve confusion and thus less of a potential to learn the material more deeply. For the induction × intervention interaction, participants were expected to perform particularly well in the Convince while Read condition when they were in either of the contradictory information conditions (True-False, False-True). This was expected because the presentation of contradictory opinions by the agents would further encourage participants to reflect on the scientific merits of the case study.

The present analyses only included those trialogue interactions in which learners engaged in the regulation task presented. Engagement in the regulation task was defined as any effort to participate in the regulation task (i.e., construct an argument with meaningful content and not a metacognitive or frozen response, open and view the explanatory text). The selection of only trialogues in which learners engaged in the regulation task was performed due to previous findings for the process and outcome of confusion regulation in the present analyses. Both the process and outcome of confusion regulation were found to be dependent upon confusion regulation effort. Thus, it was assumed that those participants who actually engaged in the regulation task to some degree would be likely to benefit from the regulation intervention. For the Convince then Read and Convince while Read conditions, the learner had to perform both tasks (construct argument, read text) to be included. Learner engagement was not predicted by induction condition, post-induction confusion judgment, or intervention condition (p's > .1). This reduced the dataset to 1059 observations. Each learning outcome measure is investigated next, followed by an investigation of the impact of confusion induction

success and regulation effort, and then the overall pattern of results for learning outcomes is discussed.

Near transfer task. Models revealed that there was not a significant induction main effect ($\chi^2(2) = .313$, $p = .855$), but there was a significant intervention main effect ($\chi^2(3) = 7.44$, $p = .059$) and significant induction × intervention interaction ($\chi^2(11) = 17.6$, $p = .091$) for the near transfer task. Table 13 shows the proportional occurrence of correctly identified flaws for the near and far transfer tasks as well as correct answers for the design-a-study task. For the intervention main effect, the Convince while Read condition was found to outperform all other conditions (Convince Only: B = .673, Read Only: B = .540, Convince then Read: B = .353). The Convince then Read condition was also found to outperform the Convince Only Condition (B = .319). Both of the main effect findings were consistent with the hypotheses for learning outcomes. The one exception was that the Read Only condition did not significantly differ from the Convince then Read condition or the Convince Only condition.

Table 13
Proportional Occurrence for Learning Outcome Dependent Measures

	Induction × Intervention Interaction								
Intervention	*Near Transfer Task*			*Far Transfer Task*			*Design A Study Task*		
Condition	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Tr	Tr-Fl	Fl-Tr
Convince Only	.370	.310	.270	.250	.250	.330	.270	.290	.210
Read Only	.330	.340	.350	.300	.400	.330	.330	.290	.360
Convince then Read	.340	.320	.470	.210	.290	.380	.220	.300	.220
Convince while Read	.500	.480	.400	.420	.400	.390	.430	.350	.320

Notes. Tr = True, Fl = False

When the significant interaction was examined, only the Convince then Read model was significant ($\chi^2(2) = 5.29$, $p = .071$), with participants performing better when

in the False-True condition compared to the True-True condition (B = .596). This finding is also consistent with hypothesized learning outcome patterns. Participants benefited from the presentation of both regulation tasks after the presentation of contradictory information. However, it was not the case that the presentation of contradictory information was particularly helpful when both regulation tasks were presented simultaneously.

False alarms were investigated for all significant models but there were no significant effects (*p's* > .1). Thus, the above results cannot be attributed to guessing.

Far transfer task. Models revealed that there was not a significant induction main effect ($\chi^2(2) = 4.27$, $p = .118$), but there was a significant intervention main effect ($\chi^2(3) = 8.71$, $p = .033$) and a significant induction × intervention interaction ($\chi^2(11) = 20.8$, $p = .036$) for the far transfer task (see Table 13). Although the induction main effect was not significant, it was approaching a marginally significant effect and showed that participants performed better when in both experimental conditions (True-False: B = .257, False-True: B = .348) compared to the no-contradiction control condition. Although this finding is not consistent with the original predictions that the presentation of contradictions alone would not increase learning, it does reveal some overall benefit to learning by the presentation of contradictory information.

For the intervention main effect, the Read Only and Convince while Read conditions performed better than the Convince Only (Read Only: B = .375, Convince while Read: B = .315) and Convince then Read conditions (Read Only: B = .611, Convince while Read: B = .551). The intervention main effect finding is both consistent and inconsistent with the hypothesized pattern. The Convince while Read condition

performed the best on the far transfer task, but the Read Only condition performed second best. Although the Read Only condition was not predicted to perform this well on the learning measures, this could be due to the positive confusion resolution outcomes previously found. The Read Only condition was found to have more resolved confusion and performed better on the Trial 3 forced-choice question than the other intervention conditions (see Confusion Regulation Outcome).

The significant interaction was examined and three models were found to be significant. The first significant model revealed the same pattern that was found for the near transfer task. When in the Convince then Read condition, participants performed better when in the False-True condition than the True-True condition ($\chi^2(2) = 6.94$, $p = .031$, B = .963). This pattern suggests that the combination of the False-True and Convince then Read conditions was particularly effective for increasing performance on transfer tasks. The remaining significant models partially supported the hypothesis that participants in the Convince while Read condition would particularly benefit from combination with the two experimental confusion induction conditions. The Convince while Read condition was found to outperform the Convince Only condition when participants were in both the True-True ($\chi^2(3) = 10.4$, $p = .016$, B = .912) and True-False conditions ($\chi^2(3) = 7.74$, $p = .051$, B = .744). When in the True-True condition, participants in the Convince while Read condition also outperformed those in the Convince then Read condition (B = 1.10). Participants in the Read Only condition benefited from the presentation of contradictory information. When in the True-False condition, participants in the Read Only condition also performed better than the Convince Only condition (B = .744). False alarms were investigated for all significant

models and the false alarm models were not significant (*p's* > .1). Thus, the results cannot be attributed to guessing.

Design-a-study task. Models revealed that there was not a significant induction main effect ($\chi^2(2) = 1.81$, $p = .404$), but there was a significant intervention main effect ($\chi^2(3) = 11.4$, $p = .010$) and a significant induction × intervention interaction ($\chi^2(11) = 17.3$, $p = .100$) for the design-a-study task (see Table 13). The intervention main effect revealed that the Convince while Read condition outperformed all other conditions (Convince Only: B = .730, Read Only: B = .458, Convince then Read: B = .645) and the Read Only condition outperformed the Convince Only condition (B = .272). Once again, the induction and intervention main effects supported the hypothesized pattern of results. The significant interaction was examined and only the True-True model was found to be significant ($\chi^2(3) = 8.01$, $p = .046$). The Convince while Read condition was found to outperform the Convince Only condition (B = .752) and Convince then Read condition (B = .966). This finding did not support the hypothesis that the Convince while Read condition would particularly benefit from the two experimental induction conditions. However, it is interesting to note that the Convince while Read condition did not outperform the Read Only condition on the design-a-study task.

Impact of confusion induction and regulation effort. Finally, the success of confusion induction and amount of regulation effort were also expected to impact learning outcomes based on previous research (D'Mello et al., 2014; Lehman et al., in preparation; Lehman et al., 2013) and the confusion resolution analyses in the present study (see Confusion Regulation Outcome). Previous research on confusion induction learning environments has shown that conditions presenting confusion-inducing stimuli

(i.e., contradictory information, false feedback) only outperformed the control condition (no-contradiction, accurate feedback) when participants were successfully confused by the confusion-inducing stimuli. Thus, the induction × confusion × intervention × regulation effort interaction was investigated next.

The prediction for these analyses was that participants would perform better on the learning measures when they were in one of the contradictory information conditions (True-False, False-True), successfully confused by the contradictory information, presented with an intervention condition that promoted engagement in the cognitive activities beneficial for learning (Convince while Read), and engaged in effortful confusion resolution. A mixed-effects logistic regression was conducted for each learning measure. Significant models were found for the near transfer ($\chi^2(47) = 78.0$, $p = .003$) and far transfer tasks ($\chi^2(47) = 61.7$, $p = .073$), but the design-a-study task was not significant ($p > .1$). Table 14 shows the proportional occurrence for each case (e.g., confused, high regulation effort, Convince while Read condition) for the near and far transfer tasks.

The overall pattern of results for the near transfer task were consistent with the previous findings in that the Convince while Read condition performed well and that the Convince then Read condition benefited from being paired with the False-True condition. Specifically, when participants were confused, had high regulation effort, and were in the True-True condition, those in the Convince while Read condition outperformed all other conditions ($\chi^2(3) = 13.4$, $p = .004$; Convince Only: B = 4.62, Read Only: B = 4.63, Convince then Read: B = 3.88). In addition, the Convince while Read condition outperformed the Convince then Read condition when participants were not confused,

had high regulation effort and were in the True-False condition ($\chi^2(3) = 10.1$, $p = .018$, B = 4.08), as did the Convince Only (B = 3.85) and Read Only conditions (B = 4.00). For the second main pattern, participants performed better when in the False-True condition than the True-True condition in the Convince then Read condition when they were confused and had low regulation effort ($\chi^2(2) = 8.32$, $p = .016$; B = 1.65). The same pattern was found when participants were not confused, had low regulation effort, and in the Read Only condition ($\chi^2(2) = 8.05$, $p = .120.018$; B = 2.67). These findings suggest that for the near transfer task the Convince while Read condition generally did the best, but the Convince then Read condition also performed well when paired with the False-True condition.

The far transfer analyses revealed a pattern of results that was consistent with the hypothesized pattern of results and also specified the circumstances under which different conditions performed well. The patterns for induction condition differences are discussed in Appendix U. The main finding was that the Convince while Read condition outperformed the Convince Only (B = 1.33) and Convince then Read conditions (B = 1.14) when participants were confused, had high regulation effort, and were in the True-False condition ($\chi^2(3) = 6.32$, $p = .097$). It is interesting to note that the Convince while Read condition did not outperform the Read Only condition in this case and although it was not a significant difference, the Read Only condition did outperform both the Convince Only and Convince then Read conditions (see Table 14). This pattern was also found when participants were confused, had high regulation effort, and were in the True-True condition, except that the Convince while Read condition also outperformed the

Read Only condition ($\chi^2(3) = 8.44$, $p = .038$; Convince Only: B = 2.13, Read Only: B = 1.69, Convince then Read: B = 1.67).

It was not the case, however, that all of the analyses revealed that the Convince while Read condition outperformed the other intervention conditions. There were two additional significant models that occurred when participants had low regulation effort and were in the True-False condition. When participants were confused, the Read Only condition outperformed the Convince then Read condition ($\chi^2(3) = 6.37$, $p = .095$; B = 1.83), whereas the Convince then Read condition outperformed all other conditions when participants were not confused ($\chi^2(3) = 18.0$, $p < .001$; Convince Only: B = -7.12, Read Only: B = -6.20, Convince while Read: B = -5.88). The distinction between when participants are confused versus not confused makes the present findings particularly interesting. Although regulation effort was low, the presence of confusion seems to have given participants extra motivation to learn the material from the text alone. However, participants who were not confused may need an additional task, such as argument construction, to trigger deeper processing and facilitate learning.

False alarms were investigated for each significant model and were found to not be significant. Therefore the findings cannot be attributed to guessing.

Discussion of learning outcome findings. As hypothesized, there were not induction condition main effects for learning, but it was the case that learners in the Convince while Read condition performed better on all three learning measures. This pattern for intervention conditions was predicted because the Convince while Read condition was expected to lead to more deliberation and problem solving as well as providing a resource to immediately address confusion and knowledge gaps. One finding

that was not expected was that the Read Only condition performed well on the far transfer task. In previous experiments (D'Mello et al., 2014; Lehman et al., in preparation; Lehman et al., 2013), simply providing an explanatory text has not been particularly effective to aid in confusion resolution.

Next, the impact of confusion induction success and regulation effort were taken into consideration. The findings for the near transfer task were not expected. In particular, most of the significant findings were when learners were not confused and the only significant finding when learners were confused occurred when learners had low regulation effort. The findings for the far transfer task were more in line with the prior predictions. In particular, the Convince while Read condition performed better than the Convince Only and Convince then Read conditions when learners were confused, had high regulation effort, and were in the True-False condition. However, it is interesting to note that under these circumstances the Convince while Read condition did not outperform the Read Only condition. Overall, the Convince while Read condition appeared to be the best for learning, however, the Read Only condition performed surprisingly well.

Table 14

Proportional Occurrence of Transfer Task Performance

	Not Confused						**Confused**					
	Low Reg. Effort			*High Reg. Effort*			*Low Reg. Effort*			*High Reg. Effort*		
	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Tr	Tr-Fl	Fl-Tr
Near Transfer Task												
Convince Only	.380	.320	.130	.500	.200	.250	.310	.320	.310	.380	.340	.300
Read Only	.170	.090	.500	.450	.250	.240	.420	.520	.240	.350	.300	.500
Convince then Read	.400	.560	.460	.370	.180	.250	.210	.230	.540	.430	.350	.520
Convince while Read	.290	.400	.290	.910	.640	.170	.560	.500	.300	.450	.410	.540
Far Transfer Task												
Convince Only	.260	.190	.400	.400	.300	.310	.240	.210	.200	.180	.290	.420
Read Only	.390	.270	.420	.450	.250	.240	.210	.450	.290	.260	.460	.390
Convince then Read	.270	.500	.460	.140	.220	.420	.180	.230	.320	.250	.270	.380
Convince while Read	.330	.100	.210	.330	.500	.170	.500	.360	.380	.480	.540	.520

Notes. Tr = True, Fl = False

5. Study 4: Motivational Confusion Interventions

Method

Participants. Participants were 180 undergraduate students from a mid-south university in the US who participated for course credit or monetary payment. Thirty-three participants received monetary payment and 147 received course credit for participation. Those who participated for monetary payment received $20 for participation. There were 112 females and 68 males in the sample. Participants' age ranged from 18 to 52 (M = 22.2, SD = 6.34). Fifty-three percent of participants were African American, 2% were Asian, 35% were Caucasian, 4% were Hispanic, and 6% were other. Prior coursework in research methods was not required for participation. Eighty-seven percent of participants had not taken a research methods course and 75% had not taken a statistics course.

Comparison to Study 3. Study 4 had an identical methodology to Study 3 with one exception, which was the nature of the confusion regulation interventions. The interventions used in Study 3 focused on pedagogical interventions, whereas the interventions used in Study 4 focus on motivational interventions. The specifics of each intervention condition are described further in the next section. After the motivational intervention, participants in all conditions were presented with an explanatory text. In addition, participants were asked to imagine that a new student had joined the conversation and this new student disagreed with them. Participants were then asked to construct an argument to convince this new student that their flaw diagnosis in the current case study is correct (identical to the Convince while Read condition in Study 3).

Confusion Regulation Intervention Manipulation. The motivation-based interventions were designed to help regulate confusion by motivating learners to persist

when confusion occurs and continue working through the short-term failure that is associated with confusion. Interventions were introduced during trialogues that identified flaws in case studies. The interventions occurred after the participant was asked to intervene (i.e., decide which agent's opinion has more scientific merit) and had made a confusion judgment (see turns 9-12 in Table 7). Table 15 shows examples of each intervention condition. The intervention conditions in Table 15 discuss the same study that was discussed in Table 7.

There were four intervention conditions. In the *General Motivational Statement* condition, participants received a supportive, encouraging statement from the tutor agent. This type of motivational intervention may help participants to regulate their confusion by providing encouragement to persist in the learning task.

In the *Material Attribute + Motivation* and *Tutor Attribute + Motivation* conditions, the tutor agent made a causal attribution statement about the source of confusion and a general motivational statement. In the *Material Attribute + Motivation* condition the tutor agent attributed any confusion to the difficulty of the material, whereas in the *Tutor Attribute + Motivation* condition confusion was attributed to an unclear explanation by the tutor agent. These two intervention conditions were similar to the interventions used in the *Affective AutoTutor* (D'Mello et al., 2011). Shifting the cause of confusion from the participant to an external source (tutor, material) was expected to further encourage participants to persist because they would feel that any confusion is not due to their own lack of knowledge or skills.

Table 15
Excerpt of Trialogue of Intervention Phase from True-False Condition

Turn	Speaker	Dialogue
		GENERAL MOTIVATIONAL STATEMENT
1	Dr. Williams	So we're getting closer, but we still haven't got this study down completely. But I know you can get it if you keep working at it! <*General motivation*>
		MATERIAL ATTRIBUTE + MOTIVATION
1	Dr. Williams	This stuff can be really challenging. I know a lot of other students have trouble with control and experimental groups. <*Attribute to material*>
2	Dr. Williams	So we're getting closer, but we still haven't got this study down completely. But I know you can get it if you keep working at it! <*General motivation*>
		TUTOR ATTRIBUTE + MOTIVATION
1	Dr. Williams	I may have not explained this very well before. I'm not always very clear when I explain control and experimental groups. <*Attribute to tutor*>
2	Dr. Williams	So we're getting closer, but we still haven't got this study down completely. But I know you can get it if you keep working at it! <*General motivation*>
		CONFUSION-SPECIFIC + MOTIVATION
1	Dr. Williams	You know, being confused is actually a good thing in learning. It means that you have an opportunity to learn about control and experimental groups really well. The best way to get past confusion is to keep trying to figure out this concept. <*Confusion-specific*>
2	Dr. Williams	So we're getting closer, but we still haven't got this study down completely. But I know you can get it if you keep working at it! <*General motivation*>
		ALL INTERVENTION CONDITIONS
3	Chris	You know what might help all of us get this stuff? Reading this chapter from my critical thinking textbook. I really think it would help. <*Introduce explanatory text*>
4	Dr. Williams	That's a great idea Chris. While we read, let's all imagine that a new person joined our conversation. Bob, this new student disagrees with you about this study. You need to put together a convincing argument to prove that you are right about the appropriateness of the control group. <*Reading purpose*>
5	Dr. Williams	Bob, type your argument to convince this new student and use the chapter to help put together your argument.<*Task instructions*>
To test this hypothesis, you need one or more comparison groups that are not exposed to the treatment… <*Explanatory text*>		
6	Bob	I think that the control and experimental groups… <*Convincing Argument*>

Finally, in the *Confusion-Specific + Motivation* condition, the tutor agent told the participant about the benefits of confusion during learning along with a general motivational statement. Although many learners feel that confusion is not a desirable state during learning (D'Mello et al., 2014), research has shown that confusion is an opportunity for learning, particularly at deeper levels (Craig et al., 2004; D'Mello et al., 2014; Graesser & D'Mello, 2012; Graesser et al., 2008; Lehman et al., in preparation; Lehman et al., 2013). This intervention was designed to reframe participant perceptions of confusion during learning. The intervention was expected to encourage participants to

persist because confusion would no longer be viewed as a negative experience (i.e., indicative of a lack of knowledge or skills) and instead would be viewed as a positive opportunity.

Design. This study had a mixed-design with confusion induction as a within-subjects factor (True-True, True-False, False-True) and confusion regulation intervention as a between-subjects factor, the same design as Study 3. The between-subjects factor had four conditions: General Motivational Statement, Material Attribute + Motivation, Tutor Attribute + Motivation, and Confusion-Specific + Motivation.

Participants completed two trialogues in each of the three confusion induction conditions with a different research methods concept discussed in each session (6 in all). Each participant completed one confusion regulation intervention condition in all six trialogues. The six research methods concepts were construct validity, control groups, experimenter bias, generalizability, random assignment, and replication. Each concept had an associated research case study that was flawed in one significant aspect (e.g., an inappropriate control group). Order of confusion induction conditions and concepts and assignment of concepts to confusion induction conditions was counterbalanced across participants with a Graeco-Latin Square. Confusion intervention condition was randomly assigned to participants.

Results and Discussion

Similar to Study 3, there were four sets of dependent measures in the analyses: confusion induction (induction phase), confusion regulation process (intervention phase), confusion regulation outcome (post-intervention phase), and learning outcome measures. A mixed-effects modeling approach was again adopted and the *lme4* package in R (Bates

& Maechler, 2010) being used to perform the requisite computations. There was one set of analyses that did not utilize a mixed-effects modeling approach (see Argument Quality Classification in Confusion Regulation Process).

Linear or logistic models were constructed on the basis of whether the dependent variable was continuous or binary, respectively. The random effects in all analyses were participant, concept, and compensation (credit, monetary). In addition, all models included order as a fixed effect (order of concept presentation). The random effects and order fixed effect were consistent across all models (control). Induction condition, intervention condition, and/or induction × intervention were the categorical fixed effect(s). The unit of analysis was the case study (or individual trialogue) for all analyses. There were 1080 observations in the present analyses.

Confusion Induction. Three mixed-effects logistic regression models were constructed to investigate induction condition differences in the induction phase dependent measures. Table 16 shows the proportional occurrence of each dependent measures as well as the coefficients from each model. Similar to Study 3, it was predicted that participants in the experimental conditions would report more confusion than the no-contradiction control condition and would respond less accurately to the forced-choice question in Trial 2, but there would be no difference for Trial 1. The model for confusion judgment was significant ($\chi^2(2) = 4.79$, $p = .091$) and revealed that participants reported more confusion when they were in both experimental conditions compared to the no-contradiction control condition. The presentation of contradictory information was again a successful method of confusion induction as in Study 3. The models for forced-choice questions in Trial 1 ($\chi^2(2) = 1.26$, $p = .533$) and Trial 2 ($\chi^2(2) = 2.97$, $p = .226$) were not

significant. The non-significant model for Trial 1 supported the hypothesized pattern and was consistent with Study 3. However, the non-significant model for Trial 2 was inconsistent with the hypothesized pattern and with Study 3, but it was the case that when participants were in both the True-False and False-True conditions they responded less accurately compared to the True-True condition.

Table 16
Proportional Occurrence of Induction Phase Dependent Measures

	Induction Condition			**Coefficients (B)**	
	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Fl	Fl-Tr
Confusion Judgment	.530	.590	.590	**.310**	**.310**
Trial 1	.700	.740	.710	.192	.045
Trial 2	.710	.660	.680	-.284	-.176

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

Confusion Regulation Process. The confusion regulation process analyses involved dependent measures that occurred during the confusion intervention manipulations. The dependent measures included argument construction + text read time (seconds), argument length (words), and argument quality. The analyses occurred over two phases. First, the quality of arguments was assessed by developing and evaluating classification models. Second, condition differences were investigated for each dependent measure.

Argument quality classification. Seven models were tested to determine which argument features were most diagnostic of argument quality. The participant arguments from Study 3 were used to develop and evaluate the classification models. The Context Model included the order of presentation, induction condition, and intervention condition.

The Induction Model included measures from the induction phase (confusion judgment, Trial 2 response time, Trial 2 response quality) and whether the participant opened the case study to review while responding to the Trial 2 forced-choice question. The Intervention Model included measures from the regulation phase (argument response time, number of words in argument, IWFWO semantic match score) and whether the participant opened the case study and the explanatory text (when applicable) to review while constructing their argument. The remaining models involved combining the features from the Context, Induction, and Intervention Models: Context + Induction, Context + Intervention, Induction + Intervention, and Context + Induction + Intervention.

Four classification algorithns from WEKA (Hall, Frank, Holmes, Pfahringer, Reutemannm & Witten, 2009) were used to build and evaluate the models: NaïveBayes, IBk (nearest neighbor with $k = 10$), j48, and LogitBoost. The majority class algorithm (ZeroR) that classifies all arguments to the most prevalent group was used as the baseline comparison. Each algorithm was evaluated using 10-fold cross-validation. Three separate classification tasks were performed. The first task consisted of making a simple correct vs. incorrect discrimination on claim quality, while the second and third tasks performed a discrimination of overall presence score (0, 1, 2) and overall quality score (high, low).

Arguments were separated into six groups based on the research methods concept. There was an average of 156 responses per group (*SD* = .894, *Range* 155 to 157). The algorithms were evaluated on each argument group for all three classification tasks. For each argument group the best algorithm (i.e., one out of the four algorithms that yielded the best performance) was selected. The best classification results were averaged across

argument groups and constituted the models. Table 17 shows the results obtained for each classification task averaged across the six groups.

Table 17
Mean (SD) of Classification Performance Across Groups

	Claim Quality		**Presence Score**		**Quality Score**	
Model	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*	*Accuracy (%)*	*Kappa*
Baseline	62.2 (9.24)	.000 (.000)	53.0 (4.88)	.000 (.000)	62.0 (7.84)	.000 (.000)
Context	64.2 (8.44)	.083 (.100)	54.9 (3.64)	.093 (.048)	63.2 (7.64)	.073 (.095)
Induction	77.4 (6.63)	.507 (.100)	52.7 (4.06)	.054 (.048)	66.3 (6.77)	.262 (.125)
Context + Induction	76.3 (7.29)	.480 (.115)	54.1 (2.87)	.105 (.043)	65.0 (6.88)	.193 (.109)
Intervention	67.2 (7.59)	.226 (.130)	70.3 (1.93)	.460 (.032)	72.1 (4.76)	.378 (.093)
Context + Intervention	67.2 (6.10)	.213 (.113)	70.3 (1.93)	.458 (.028)	71.4 (5.52)	.345 (.113)
Induction + Intervention	76.9 (6.34)	.492 (.096)	68.0 (2.22)	.417 (.031)	75.0 (2.91)	.444 (.065)
Context + Induction + Intervention	74.2 (6.57)	.427 (.111)	68.6 (1.44)	.429 (.034)	75.4 (3.15)	.454 (.079)

The Induction Model was most successful for discriminating between correct vs. incorrect claims in the argument. The Induction Model performed significantly better than the Baseline Model, $t(5) = 5.98$, $p = .002$; and performed significantly better than all other models except for the Context + Induction Model, $t(5) = 1.55$, $p = .181$; and Induction + Intervention Model, $t(5) = .514$, $p = .629$. The Intervention Model was most successful for discriminating between presence scores of 0 (no claim or evidence), 1 (claim or evidence), and 2 (claim and evidence). The Regulation Model performed significantly better than the Baseline Model, $t(5) = 9.50$, $p < .001$; and performed significantly better than all other models except for the Context + Intervention Model, $t(5) = .006$, $p = .995$; and Induction + Intervention Model, $t(5) = 1.65$, $p = .159$. Finally

the full model of Context + Induction + Intervention performed the best at discriminating between low and high overall quality scores. The Context + Induction + Intervention performed significantly better than the Baseline Model, $t(5) = 4.74$, $p = .005$; and outperformed all other models except for the Induction + Intervention Model, $t(5) = .455$, $p = .668$. The Induction, Intervention, and Context + Induction + Intervention Models were then used to classify participant arguments in Study 4 for claim quality, overall presence score, and overall quality score, respectively.

Analyses. Three mixed-effects linear regressions were constructed for each dependent measure: induction main effect, intervention main effect, and induction × intervention interaction. Similar to Study 3, the contradictory information conditions were expected to trigger greater processing time than the no-contradiction control. This hypothesis was not confirmed by the present analyses. The induction main effect models were not significant for any of the dependent measures (p's > .1). However, this hypothesis was indirectly supported by the finding that confusion was a significant predictor of the total time to read the explanatory text and construct an argument ($F(1,1080) = 8.48$, $p = .004$). Confused participants had longer overall times than not confused, similar to the finding in Study 3.

The following pattern was expected for all dependent measures for the intervention conditions: *General Motivational Statement* < *Material Attribute + Motivation* < *Tutor Attribute + Motivation* < *Confusion-Specific + Motivation*. This pattern was expected because of the impact the intervention would have on participants' causal attributions, based on attribution theory (Batson et al., 1995; Heider, 1958; Weiner, 1986).

The *General Motivational Statement* condition was expected to motivate participants to persist and put forth effort to resolve their confusion the least because it did not address the causal attribution for confusion. Both of the attribute + motivation conditions (material, tutor) were expected to motivate participants more than the *General Motivational Statement* condition because they addressed the causal attribution for confusion and removed responsibility from the participant. The *Material Attribute + Motivation* condition was expected to motivate participants less than the *Tutor Attribute + Motivation* condition because of the stability aspect of each type of attribution. Both attributions were external and uncontrollable in that the participant cannot impact the difficulty of the current concept or the quality of the tutor agent's explanation. However, the difficulty of the concept may be viewed as more stable than the unclear explanation. In other words, if the concept was very difficult as the tutor agent asserted, then increased effort on the part of the participant (i.e., reading the text more deeply) may not lead to a change in outcome (i.e., confusion resolution). Alternatively, an increase in effort could potentially overcome the tutor agent's unclear explanation and lead to a change in outcome.

Finally, the *Confusion-Specific + Motivation* condition was expected to motivate participants the most to persist because it engaged in a form of attributional retraining. The two attribute + motivation conditions shifted the attribution from an internal to an external source, or reaffirmed an external source; however, confusion would still be perceived as a negative experience in both of these conditions. In the *Confusion-Specific + Motivation* condition, on the other hand, confusion was reframed as a beneficial experience for learning. This shift from a negative to a positive experience was expected

to increase participants' motivation to persist and put in the effort to resolve their confusion.

The hypothesized pattern for the intervention condition main effect was also not supported in the present analyses. The models for the intervention main effect and the induction × intervention interaction were not significant for any of the dependent measures (*p's* > .1).

Argument quality was further examined by investigating the induction × confusion (confused, not confused) × intervention × regulation effort (high low) as was done in Study 3. Table 18 shows the proportional occurrence for each argument quality dependent measure. The interaction was significant for the presence score of 1 ($\chi^2(47) = 80.9$, $p = .002$) and 2 ($\chi^2(47) = 82.2$, $p = .001$), but not for the overall score ($p > .1$). A presence score of 1 represented an argument that contained either a claim or evidence, whereas a presence score of 2 represented an argument that contained both a claim and evidence. It is important to note, however, that neither score is dependent upon the quality of the claim or evidence.

The critical case in these analyses was when participants were confused, had low regulation effort, and were in the Material Attribute + Motivation condition. Participants were less likely to have a presence score of 1 when in the True-False condition compared to the True-True condition ($\chi^2(2) = 5.00$, $p = .082$, B = 1.76), but more likely to have a presence score of 2 ($\chi^2(2) = 6.69$, $p = .035$, B = 2.18). Participants were also more likely to have a presence score of 2 when in the False-True condition (B = 1.69). This finding

Table 18

Proportional Occurrence of Argument Quality Dependent Measures

	True-True				True-False				False-True			
	GMS	*MA*	*TA*	*CS*	*GMS*	*MA*	*TA*	*CS*	*GMS*	*MA*	*TA*	*CS*
Presence Score												
Zero												
NC-LR	.000	.080	.000	.000	.030	.000	.000	.050	.000	.050	.070	.000
NC-HR	.000	.000	.000	.000	.000	.000	.000	.000	.050	.000	.000	.070
C-LR	.030	.080	.000	.000	.040	.040	.080	.100	.000	.040	.000	.000
C-HR	.000	.040	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
One												
NC-LR	.350	.540	.310	.350	.450	.400	.570	.530	.320	.320	.360	.330
NC-HR	.160	.130	.220	.280	.080	.210	.110	.310	.050	.180	.310	.070
C-LR	.470	.580	.500	.440	.360	.370	.310	.380	.480	.420	.330	.430
C-HR	.320	.110	.140	.140	.140	.180	.040	.150	.160	.160	.140	.290
Two												
NC-LR	.650	.380	.690	.650	.520	.600	.430	.420	.680	.630	.570	.670
NC-HR	.840	.870	.780	.720	.920	.790	.890	.690	.900	.820	.690	.860
C-LR	.500	.350	.500	.560	.610	.590	.620	.520	.520	.540	.670	.570
C-HR	.680	.850	.860	.860	.860	.820	.960	.850	.840	.840	.860	.710
Claim Quality												
NC-LR	.700	.540	.620	.600	.520	.400	.650	.530	.730	.580	.540	.670
NC-HR	.680	.710	.740	.800	.330	.430	.610	.560	.620	.550	.380	.710
C-LR	.690	.620	.610	.750	.540	.700	.620	.570	.560	.620	.670	.650
C-HR	.680	.480	.590	.620	.860	.710	.610	.740	.680	.710	.450	.630
Quality Score												
NC-LR	.600	.150	.420	.300	.380	.330	.350	.260	.320	.320	.390	.400
NC-HR	.680	.750	.570	.520	.670	.430	.390	.500	.620	.450	.460	.430
C-LR	.280	.350	.280	.130	.290	.410	.420	.240	.480	.460	.330	.350
C-HR	.580	.520	.410	.480	.670	.590	.520	.620	.420	.650	.480	.390
Semantic Match Score												
NC-LR	.301	.211	.405	.266	.319	.423	.262	.438	.375	.333	.394	.348
NC-HR	.287	.420	.466	.316	.476	.321	.415	.272	.319	.400	.391	.337
C-LR	.392	.312	.262	.281	.373	.345	.349	.280	.333	.357	.415	.255
C-HR	.410	.371	.510	.374	.300	.450	.322	.392	.395	.395	.430	.353

Notes. GMS = General Motivational Statement, MA = Material Attribute + Motivation, TA = Tutor Attribute + Motivation, CS = Confusion-Specific, NC = not confused, C = confused, LR = low regulation, HR = high regulation

neither supports nor refutes the proposed pattern of findings. Given the nature of the interventions, it would be expected that high regulation effort would be a key component to success at any point in the learning process (i.e., regulation process, regulation outcome, learning outcome). The present findings, however, show that participants generated higher quality arguments when in the experimental induction conditions when they were successfully confused, but put in low regulation effort. There was an exception to this pattern. When participants were confused, had high regulation effort, and were in the General Motivational Statement condition, they were more likely to have a presence score of 2 when in the True-False condition compared to the True-True condition ($\chi^2(2) = 7.78$, $p = .022$, B = 6.92). Although it was unexpected that the General Motivational Statement condition would generate higher quality arguments, it seems plausible that it could happen when participants are successfully confused and put in the effort to resolve their confusion.

Confusion Regulation Outcome. The confusion regulation outcome dependent measures consisted of confusion resolution outcome and response quality on the forced-choice question in Trial 3. Confusion resolution was assessed in the same manner as in Study 3: none (N = 368), resolved (N = 356), unresolved (N = 258), and created (N = 91). Overall, participants in the most motivational condition (Confusion-Specific + Motivation) were expected to resolve their confusion and respond correctly to the Trial 3 forced-choice question more than participants in the other conditions. Mixed-effects logistic regressions were conducted and did not support this hypothesized pattern of findings for the Trial 3 forced-choice question. There were no significant differences found for performance on the Trial 3 forced-choice questions (*p's* > .1). The induction ×

intervention interaction was also not significant for the confusion regulation outcomes. Table 19 shows the proportional occurrence for each dependent measure.

Table 19
Proportional Occurrence of Post-Regulation Phase Dependent Measures

	Condition						
	Tr-Tr	*Tr-Fl*	*Fl-Tr*	*GMS*	*MA*	*TA*	*CS*
Confusion Resolution							
None	.390	.320	.320	.390	.270	.400	.310
Resolved	.300	.350	.350	.300	.340	.320	.370
Unresolved	.230	.250	.250	.240	.300	.190	.230
Created	.080	.080	.090	.070	.090	.090	.090
Trial 3	.140	.140	.180	.170	.140	.150	.150
Low Regulation Effort							
None	.400	.370	.420				
Resolved	.290	.280	.290				
Unresolved	.250	.270	.240				
Created	.060	.090	.060				
High Regulation Effort							
None	.380	.270	.220				
Resolved	.300	.420	.410				
Unresolved	.210	.230	.260				
Created	.110	.080	.120				

Notes. Tr = True, Fl= False, GMS = General Motivational Statement, MA = Material Attribute + Motivation, TA = Tutor Attribute + Motivation, CS = Confusion-Specific + Motivation

Confusion resolution was investigated and there were only significant models for no confusion. The same pattern that was found in Study 3 was found again. Participants in both experimental conditions were less likely to have no confusion than the no-contradiction control condition ($\chi^2(2) = 6.10$, $p = .047$, True-False: B = .365, False-True: B = .395). Although the model was not significant, the proportional occurrence of resolved confusion was investigated for each induction condition. It was the case that both experimental conditions (True-False: .350, False-True: .350) were more likely to have resolved confusion than the no-contradiction control (.300). It was also the case that

both experimental conditions had higher proportional occurrences of resolved confusion than unresolved confusion (True-False: .250, False-True: .250). So it seems that there is a general pattern revealing that participants were able to more successfully resolve their confusion when in the contradictory information conditions.

There was also an intervention main effect for the no confusion outcome; however, it did not support the current hypothesis ($\chi^2(3) = 8.90$, $p = .031$). The General Motivational Statement (B = .712) and Tutor Attribute + Motivation conditions (B = .766) were more likely to have no confusion than the Material Attribute + Motivation condition. In addition, the Tutor Attribute + Motivation condition was also more likely to have no confusion than the Confusion-Specific condition (B = .456). Once again, the proportional occurrence of resolved confusion was investigated to determine the overall pattern, which was General Motivational Statement (.300) < (Tutor Attribute + Motivation (.320) = Material Attribute + Motivation (.340)) < Confusion-Specific + Motivation (.370). Although the model was not significant, this general pattern follows the prediction made based on attribution theory (Batson et al., 1995; Heider, 1958; Weiner, 1986).

Next, the impact of regulation effort on confusion resolution was investigated as in Study 3. The induction × intervention × regulation effort (high, low) interaction was not significant for any of the confusion resolution outcomes (*p's* > 1). However, the induction × regulation effort interaction was significant for no confusion ($\chi^2(5) = 18.4$, $p = .002$) and resolved confusion ($\chi^2(5) = 14.1$, $p = .015$). Table 19 shows the proportional occurrence for each confusion resolution outcome when split by regulation effort. The interactions were then examined and revealed that models for low regulation effort were

not significant for no confusion or resolved confusion (*p's* > .1), but the models for high regulation effort were significant for both confusion resolution outcomes.

Similar to the induction condition main effect, participants were found to have less none (i.e., no confusion at T1 or T2) when in both experimental conditions compared to the no-contradiction control condition ($\chi^2(2) = 12.4$, $p = .002$, True-False: B = .575, False-True: B = .879). However, it was also the case that participants were more likely to have resolved confusion when in both experimental conditions compared to the no contradiction control ($\chi^2(2) = 7.71$, $p = .021$, True-False: B = .612, False-True: B = .525). It appears that the general but non-significant pattern found in the induction main effect analyses has again occurred in the high regulation cases. This finding was similar to Study 3 in that participants were more likely to resolve their confusion when they put in the necessary effort.

Learning Outcomes. Performance on the near transfer, far transfer, and design-a-study tasks was next investigated. As in Study 3, performance on the transfer tasks was assessed by hits (i.e., correctly identifying a flaw) and false alarms (i.e., incorrectly identifying a flaw), whereas design-a-study task performance was assessed by selecting the correct multiple-choice response. Three logistic regression models were constructed to investigate differences based on induction and intervention condition for each learning measure. Similar to Study 3, only those cases in which participants engaged in the regulation task (i.e., opened the explanatory text and provided an argument with meaningful content) were included in the present analyses. This reduced the dataset to 913 observations.

Induction and intervention condition differences. The induction and intervention condition main effects were investigated first. Table 20 shows the proportional occurrence of correct responses for each posttest. The performance for each induction and intervention condition are displayed. The induction main effect model was expected to not be significant for all three posttests as in Study 3. This hypothesis was supported for the transfer tasks, but not for the design-a-study tasks. The induction main effect models were not significant for the transfer tasks (*p's* > .1), but it was for the design-a-study task and revealed that when participants were in both experimental conditions they performed worse than the no-contradiction control condition ($\chi^2(2) = 11.9$, $p = .003$, True-False: B = .324, False-True: B = .640). This finding may suggest that the design-a-study task did not benefit from the process of inducing confusion and then resolving it.

Table 20
Proportional Occurrence of Learning Measures

	Condition						
	Tr-Tr	*Tr-Fl*	*Fl-Tr*	*GMS*	*MA*	*TA*	*CS*
Near Transfer Task	.360	.340	.350	.320	.380	.330	.370
Far Transfer Task	.360	.340	.360	.390	.290	.340	.380
Design-A-Study Task	.350	.280	.240	.300	.300	.290	.270

Notes. Tr = True, Fl = False, GMS = General Motivational Statement, MA = Material Attribute + Motivation, TA = Tutor Attribute + Motivation, CS = Confusion-Specific + Motivation

The expected pattern of performance was General Motivational Statement < Material Attribute + Motivation < Tutor Attribute + Motivation < Confusion-Specific + Motivation. Similar to Study 3, this pattern was expected because of the potential for successful confusion resolution in each condition. As previously mentioned (see Confusion Regulation Outcome), participants in the Confusion-Specific + Motivation

condition were expected to perform better than those in all other intervention conditions because they would be motivated to persist and put in the effort needed to successfully resolve confusion through attributional retraining. The distinction between the Material Attribute + Motivation and Tutor Attribute + Motivation conditions was due to participant perceptions about the stability of each outcome. The unclear tutor explanation attribution could be viewed as less stable than the difficult material attribution and thus be more motivating for participants to change their outcome (i.e., confusion) through effortful cognitive activities. Finally, participants in the General Motivational Statement were not expected to perform well on the learning measures because the causal attribution associated with the experience of confusion was not addressed.

The intervention main effect hypothesis was not supported by the current analyses. The intervention main effect model was not significant for any of the posttests (*p's* > .1). Although the models were not significant, the proportional occurrence of correct responses was still investigated for each posttest. The Confusion Specific condition was found to be functionally equivalent to the top scoring condition (i.e., .01 difference in scores) in both transfer tasks. The general pattern then shows that the Confusion-Specific intervention was somewhat effective at promoting learning through attributional retraining. However, all other intervention conditions outperformed the Confusion Specific + Motivation condition on the design-a-study task.

The expected pattern for the induction × intervention interaction was similar to that in Study 3. When participants were in the contradictory information conditions (True-False, False-True), they were expected to perform particularly well in the Confusion-Specific + Motivation condition on the posttest. This was expected because

the posed contradiction would also encourage participants to engage in the beneficial cognitive activities needed for confusion resolution (e.g., reflection, deliberation). This hypothesis was also not confirmed due to the fact that the induction × intervention interaction was not significant for any of the posttests (*p's* > .1). It was the case, however, that participants performed better on all three learning measures when they had high regulation effort (Near Transfer Task: $\chi^2(1) = 3.37$, $p = .066$; Far Transfer Task: $\chi^2(1) = 4.86$, $p = .028$; Design-A-Study Task: $\chi^2(1) = 7.71$, $p = .005$). This finding suggests that when participants were properly motivated through whatever means (internal, external), they were able to perform well on all of the posttests.

Impact of confusion induction and regulation effort. Next, the impact of confusion induction success and regulation effort were investigated. The induction × confusion × intervention × regulation effort interaction was tested with a mixed-effects logistic regression model for each learning measure and was significant for the near transfer ($\chi^2(47) = 69.3$, $p = .019$) and design-a-study tasks ($\chi^2(47) = 71.3$, $p = .013$), but not for the far transfer task ($p > .1$). The significant interactions were further examined by dividing the data into separate groups for each case (e.g., not confused, low regulation effort, general motivational statement condition). Table 21 shows the proportional occurrence of correct responses for the near transfer and design-a-study tasks. Intervention condition differences are discussed next and induction condition differences can be found in Appendix V.

The experimental induction conditions were found to be the only significant models when individual cases were investigated for the near transfer task. Overall, the

Table 21

Proportional Occurrence of Near Transfer and Design-A-Study Tasks

	Not Confused						**Confused**					
	Low Reg. Effort			*High Reg. Effort*			*Low Reg. Effort*			*High Reg. Effort*		
	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Tr	Tr-Fl	Fl-Tr	Tr-Tr	Tr-Fl	Fl-Tr
Near Transfer Task												
GMS	.400	.100	.320	.370	.500	.330	.410	.250	.370	.320	.480	.160
MA	.380	.400	.260	.460	.570	.640	.230	.330	.380	.370	.320	.420
TA	.270	.220	.390	.350	.500	.230	.170	.230	.220	.450	.220	.520
CS	.250	.530	.200	.360	.440	.140	.310	.430	.220	.480	.380	.450
Design A Study Task												
GMS	.500	.170	.320	.370	.250	.240	.280	.320	.300	.370	.290	.210
MA	.380	.130	.210	.290	.360	.270	.310	.330	.150	.590	.210	.350
TA	.380	.220	.360	.480	.330	.380	.170	.350	.220	.180	.350	.100
CS	.250	.160	.270	.520	.310	.140	.130	.240	.170	.340	.380	.180

Notes. Tr = True, Fl = False, GMS = General Motivational Statement, MA = Material Attribute + Motivation, TA = Tutor Attribute + Motivation, CS = Confusion-Specific + Motivation

models revealed that the Confusion-Specific + Motivation condition was not superior to all conditions, but did outperform other conditions on the near transfer task. Specifically, the Confusion-Specific + Motivation condition outperformed both the General Motivational Statement (B = 2.41) and Tutor Attribute + Motivation conditions (B = 1.42) when participants were not confused, had low regulation effort, and were in the True-False condition ($\chi^2(3) = 11.7$, $p = .008$). The Material Attribute + Motivation condition was also found to outperform the General Motivational Statement condition in this case (B = 1.83). Although this pattern supports the hypothesis that the Confusion Specific + Motivation condition would perform the best on learning measures, it is somewhat counterintuitive. Participants were not confused and did not put forth a great amount of effort, most likely because they did not have any confusion to resolve, so it is anomalous that an intervention specifically targeting confusion would be helpful in these circumstances.

There was a more intuitive finding when participants were confused, had high regulation effort, and were in the False-True condition. In this case the General Motivational Statement condition was outperformed by all other intervention conditions ($\chi^2(3) = 7.85$, $p = .049$, Material Attribute + Motivation: B = 1.53; Tutor Attribute + Motivation: B = 1.93, Confusion Specific + Motivation: B = 1.51). This finding is more intuitive because participants were confused, so interventions that targeted the causal attributions for confusion was context appropriate. In addition, these participants put in greater effort during the confusion regulation task. The end result of increased learning on the near transfer task easily follows from the series of events. However, this finding is not entirely consistent with the hypothesized pattern because the attributional retraining

strategy employed in the Confusion Specific + Motivation condition did not facilitate greater learning than the attributional shift strategy used in the Material Attribute + Motivation and Tutor Attribute + Motivation conditions.

There was one instance in which the Confusion Specific + Motivation condition was outperformed on the near transfer task. When participants were not confused, had high regulation effort, and were in the False-True condition, those in the Material Attribute + Motivation condition outperformed both the Confusion Specific + Motivation (B = 2.46) and Tutor Attribute + Motivation conditions (B = 1.99, $\chi^2(3) = 6.97$, $p = .073$). False alarms were investigated for each significant model and were found to not be significant. Therefore the results cannot be attributed to guessing.

The significant models for the design-a-study task all involved the cases in which participants were confused and had high regulation effort. When participants were in the True-True condition, the Tutor Attribute + Motivation condition was outperformed by all intervention conditions (GMS: B = 1.36, MA: B = 2.28, CS: B = 1.23, $\chi^2(3) = 9.66$, $p = .022$). In addition, the Material Attribute + Motivation condition also outperformed the Confusion Specific + Motivation condition (B = 1.05). This pattern may have occurred because of the combination of the induction and intervention condition. In the True-True condition the agents agree and present a correct opinion, but then the tutor agent states that she does not explain this topic well in the Tutor Attribute + Motivation condition. This juxtaposition of agreement which creates some degree of certainty with explanation is due to the tutor agent explaining the concept poorly may cause participants to question the correct response and believe that the incorrect response is correct during the trialogue. The difference between the Material Attribute + Motivation and Confusion Specific +

Motivation conditions is particularly interesting when considered in the context that the opposite pattern was found when learners were in the True-False condition ($\chi^2(3) = 4.63$, $p = .100$). Specifically, the Confusion Specific + Motivation condition outperformed the Material Attribute + Motivation condition on the design-a-study task (B = 1.27). Based on these findings it appears that the presentation of contradictory information was an important factor in the effectiveness of the Confusion Specific + Motivation condition. It may be the case that confusion triggered in the True-True condition is different from confusion triggered by the presentation of contradictory information by the two agents. For example, the True-True condition confusion could be more similar to hopeless confusion since the participant is confused even when it would be possible to just adopt the opinion being proposed by both of the agents.

Discussion of learning outcome findings. When learning outcomes were investigated, there was not a significant induction main effect as predicted and as was found in Study 3. The intervention condition hypothesis that the Confusion-Specific + Motivation condition would perform the best on the learning measures was found for the transfer tasks, although the models were not significant. However, there were significant models revealing that learners who put in more regulation effort were able to perform better on all three learning measures. The combination of regulation effort and confusion induction success were then investigated and there was some evidence that addressing attributions was more helpful than a general motivational statement. However, the findings were not completely consistent with the hypothesized pattern. The only finding that was consistent with the predicted patterns was that when learners were confused, had high regulation effort, and were in the True-False condition those in the Confusion-

Specific + Motivation condition performed better on the design-a-study task than those in the Material Attribute + Motivation condition. Interestingly, it was also the instances in which learners were confused, had high regulation effort, and were in the True-False condition that the predicted pattern was found in Study 3. It may be that case that more information about the learner is needed when attributions are being addressed by an intervention. For example, if learners are making the attribution that confusion is due to an internal source and is permanent, it may be more difficult to convince them that confusion is a learning opportunity than learners who attribute confusion to a different source.

6. General Discussion

Overview of Research

The present dissertation adopted a multi-pronged approach to investigate interventions to regulate confusion during learning. Confusion has been found to frequently occur during learning (Craig et al., 2004; D'Mello & Graesser, 2011; Graesser et al., 2007; Lehman et al., 2008) and can be beneficial for learning, particularly at deeper levels (Craig et al., 2004; D'Mello & Graesser, 2011; D'Mello et al., 2014; Graesser et al., 2007; Lehman et al., in preparation; Lehman et al., 2013). However, it is not the case that all learners are able to resolve their confusion and reach a deeper level of understanding (D'Mello & Graesser, 2012b). The inability to resolve confusion may be due to a lack of knowledge, skill, or effort. Thus, it is important to investigate confusion interventions that address these different factors that can contribute to hopeless confusion. The present dissertation investigated confusion regulation interventions in three contexts. First, learner preferences for confusion regulation interventions were investigated (Study 1). Second, the way in which expert human tutors handled instances of learner confusion were investigated (Study 2). Third, confusion regulation interventions were directly evaluated in a learning environment that experimentally induced confusion via the presentation of contradictory information. In this learning environment, pedagogical (Study 3) and motivational interventions (Study 4) to aid confusion resolution were investigated.

Learner preferences were investigated with an online survey study. Learners rated which interventions they felt would help them overcome their confusion. Learners were found to prefer an intervention that would help them solve their current confusion (e.g.,

more information, feedback, correct answer) as opposed to supportive comments or a change of task, related or unrelated. Interventions were also rated for how they would help learners overcome boredom and frustration. The findings revealed that learners preferred different interventions for boredom, confusion, and frustration. Thus, it does not appear that learners perceive all negatively-valenced emotions during learning as similar.

Next, expert human tutor responses to learner confusion were investigated by analyzing the dialogue moves that occurred following instances of learner confusion. Generally, tutors adopted a more pedagogical approach to handling learner confusion. In particular, tutors provided direct instruction and explanation after learner confusion. Interestingly, this response to learner confusion is consistent with learner preferences for confusion interventions. In addition to learners preferring more information and tutors providing direction instruction after confusion, tutors also provide the correct answer frequently after learner confusion and learners rated receiving the correct answer as helpful to resolve their confusion. The apparent approach adopted by tutors was most consistent with the strategy proposed by Vygotskian theory (1978) in that tutors were more directly helping learners to resolve their confusion and not posing questions to the learners (VanLehn et al., 2003) or providing motivational support during this struggle (Lepper & Woolverton, 2002). Tutors also handled confusion differently than other emotions (i.e., anxiety, frustration, and happiness), which is another similarity with the results from the learner preferences study. In addition, tutors did not handle confusion the same as learner questions and incorrect answers. Incorrect answers, in particular, have been used as a model to address impasses (VanLehn et al., 2003) and uncertainty

(Forbes-Riley & Litman, 2011). Finally, tutors did not handle all instances of confusion as the same. The cognitive state (e.g., incorrect answer, question, metacognitive statement) that occurred with confusion influenced tutorial dialogue.

Finally, interventions to regulate confusion were investigated within a learning environment that induced confusion. Previous research on interventions to regulate confusion (D'Mello et al., 2010) and uncertainty (Forbes-Riley & Litman, 2011) have used natural occurrences of confusion and uncertainty during interactions with an intelligent tutoring system. In both studies, learning benefits from the affect intervention were only beneficial for some learners. Learners with lower prior knowledge, who may have been more likely to experience confusion during learning, benefited the most from Affective AutoTutor (D'Mello et al., 2010) and learners who had more experiences of uncertainty benefited the most from UNC-ITSpoke (Forbes-Riley & Litman, 2011). Thus, it is difficult to determine the effectiveness of each intervention when some learners may have received the intervention multiple times and other learners may have never actually received the intervention. The present dissertation addressed this issue by evaluating pedagogical and motivational interventions within an environment that experimentally induced confusion. In both studies confusion was successfully induced by having the two animated pedagogical agents present contradictory information while discussing the scientific merits of research case studies.

Both pedagogical and motivational interventions were investigated with respect to how they influenced confusion regulation (i.e., confusion resolution) and learning outcomes. Table 22 shows the main results from Studies 3 (pedagogical) and 4 (motivational). For each result the hypothesized pattern is displayed for both the

induction and intervention main effects, where applicable, as well as the pattern found in each study. In addition, non-parametric tests (signed-rank test, Mann-Whitney U test, Kruskal-Wallis test) were conducted to compare the induction and intervention conditions without taking into consideration adjustments due to random effects.

The main findings from Studies 3 and 4 encompassed confusion induction, confusion regulation process, confusion regulation outcome, and learning outcomes. For confusion induction, the presentation of contradictory information successfully induced confusion in both Studies 3 and 4. Although more confusion was reported in the two experimental confusion induction conditions, it was not the case that these two conditions had increased processing time during the confusion regulation process, as hypothesized. However, this pattern was still "partially" supported due to the fact that overall when learners reported confusion after the presentation of contradictory information they had longer processing time than when they did not report confusion in both Studies 3 and 4. The predicted patterns for intervention main effects for both regulation process time and argument quality were not found for either study, with one exception (see Table 22). Overall the findings for regulation process time and argument quality were mixed and did not present clear and consistent effects.

The pattern for confusion resolution outcome was also somewhat unclear for both Studies 3 and 4. However, it was the case that learners who put in more effort during the regulation task were more likely to resolve their confusion. This is consistent with impasse driven theories of learning that predict impasses will be resolved, and learning is likely to occur, when learners engage in beneficial, *effortful* cognitive activities such as deliberation, problem solving, and reflection (Brown & VanLehn, 1980; VanLehn et al.,

2003). Given this finding it was necessary to consider both regulation effort and confusion induction success when investigating learning outcomes.

Finally, learning outcomes were investigated for both Studies 3 and 4. The findings from the pedagogical intervention study were generally consistent with the prediction that the condition that most encouraged learners to deliberate between the competing perspectives and provided resources to address knowledge gaps would be the most helpful (i.e., Convince while Read condition). In particular, the Convince while Read condition was found to be effective when learners were confused, had high regulation effort, and were in a condition that presented contradictory information. Similar to the confusion resolution findings, the fact that the Convince while Read condition was effective, and particularly when learners were confused and put in more effort on the regulation task, is consistent with cognitive disequilibrium (Festinger, 1957; Graesser et al., 2005; Piaget, 1952) and impasse-driven theories of learning (Brown & VanLehn, 1980; VanLehn et al., 2003). There were two unexpected findings in terms of learning outcomes. First, the Convince then Read condition was less effective than was expected. However, the Convince then Read condition was effective when learners put in low effort and were in a contradictory information condition. This finding is somewhat perplexing because learners did not put in the effort to resolve their confusion, but were still able to perform well on the posttest. Second, the Read Only condition had learning outcomes similar to the Convince while Read condition in many instances. This may have been due to the fact that the Read Only condition was found to have more successful confusion resolution, which would also support cognitive disequilibrium (Festinger, 1957; Graesser et al., 2005; Piaget, 1952) and impasse-driven theories of learning

Table 22
Summary of Main Findings from Studies 3 and 4

	Study	Pattern		Test	Supported?
		Predicted	Observed		
Confusion Induction	S3 & S4	S3 & S4: TF, FT > TT	TF, FT > TT	$p = .001$, $p = .003$ $p = .071$, $p = .057$	Yes Yes
Regulation Process					
Processing Time	S3 & S4	S3 & S4: TF, FT > TT	TF, FT = TT	$p's > .100$, $p's > .100$ $p = .803$, $p = .562$	Partially Partially
	S3	S3: CtR > RO	CtR = RO	$p = .805$	No
		S3: CO = CtR	CO = CtR	$p = .314$	Yes
	S4	CS > (TA = MA) > GMS	CS = TA = MA = GMS	$p = .182$	No
Argument Quality	S3	CwR > CO = CtR	CO = CtR = CwR	$p = .546$	No
	S4	CS > (TA = MA) > GMS	CS = TA = MA = GMS	$p = .485$	No
Confusion Resolution	S3	CwR > CtR > RO > CO	RO > CtR > CwR > CO*	$p = .398$	No
	S4	CS > (TA = MA) > GMS	CS > (TA = MA) > GMS*	$p = .678$	Partially
Learning Outcomes					
Near Transfer Task	S3	CwR > CtR > RO > CO	CwR > CO, RO, CtR | CtR > CO	$p = .059$	Yes
	S4	CS > (TA = MA) > GMS	CS = TA = MA = GMS	$p = .614$	Partially
Far Transfer Task	S3	CwR > CtR > RO > CO	RO = CwR > CO = CtR	$p = .026$	Partially
	S4	CS > (TA = MA) > GMS	CS = TA = MA = GMS	$p = .091$	Partially
Design-A-Study Task	S3	CwR > CtR > RO > CO	CwR > CO, RO, CtR | RO > CO	$p = .016$	Partially
	S4	CS > (TA = MA) > GMS	CS = TA = MA = GMS	$p = .686$	No

Notes. * = general pattern was observed, but was not significant in previous analyses

(Brown & VanLehn, 1980; VanLehn et al., 2003).

The findings from the motivational interventions study were overall less clear than the pedagogical study. However, there was evidence that interventions that addressed learner attributions were more effective than those only addressing general motivation. It may be the case that a larger intervention is needed when attempting to alter learner attributions. For example, Perry et al. (2010) used training sessions that involved watch a short video and discussion to retrain learners to perceive effort as the key component to academic success, as opposed to intelligence. Interventions to convince learners that confusion is actually a beneficial state for learning may take more than a couple of sentences stated by the tutor agent while learning is occurring. As noted before, it may also be necessary to assess learners' perceptions of confusion generally and their current attribution for confusion in the specific trialogue. These two pieces of information will undoubtedly aid in selecting an intervention that positively impacts attributions and encourages persistence in the face of confusion and struggles.

Overall, the present dissertation investigated interventions to regulate confusion during learning. In other words, how can interventions be deployed to keep learners in the virtuous affective cycle and avoid the vicious cycle identified by D'Mello and Graesser (2012b) (see Figure 1). Across the four studies that were discussed, there appears to be a commonality in that more information is helpful when learners are confused. Learners preferred more information when confused, expert tutors delivered more information after learner confusion, and learners benefited most from interventions that supplied additional information during interactions with animated pedagogical agents. However, it

also appears that the way in which that information is presented is important to insure that learners make use of that information.

The issue of how to present additional information is critical to developing interventions that facilitate confusion resolution and deeper learning. Expert human tutors provide direct instruction most frequently after learner confusion; however, they are also breaking down problems into more manageable sub-problems, asking follow-up questions, and employing motivational dialogue moves. This suggests that expert tutors view it as necessary to couch the additional information provided in direct instruction within a context that requires learners to persist with the current problem or with motivational statements that encourage learners to persist through their confusion. This finding is consistent with the uncertainty-adaptive UNC-ITSpoke (Forbes-Riley & Litman, 2011) and the results from Studies 3 and 4 in the present dissertation. However, it was not the case that any of the strategies deployed in conjunction with additional information were effective for all learners.

The way in which expert human tutors responded to learner confusion may help to explain why the strategies used in UNC-ITSpoke (Forbes-Riley & Litman, 2011) and Studies 3 and 4 were not effective for all learners. The tutor-learner pairs in Study 2 were already working together prior to the study in which this data was collected. The tutors then had an understanding of not only the learner's abilities but also his or her perceptions of the topic being tutored and learning more generally as well as his or her response to challenges and academic failures. In other words, expert tutors were likely responding to both learners' emotions and their individual characteristics. It could also be the case that at different points in the tutoring session (e.g., beginning vs. end) tutors

deploy different strategies or tutors may shift strategy depending on the topic or problem currently being discussed (e.g., easy or difficult topic, topic a learner does or does not like). The variety in expert tutor responses may reveal that the most effective method to facilitate confusion resolution must adapt to both the learner and the current learning context.

The issue of adapting to the individual learner's characteristics and the individual tutoring session characteristics is also relevant to the application of the present confusion regulation interventions in other learning contexts. How would these interventions function in a small group setting (computer-mediated or face-to-face) or in a classroom setting? It may be the case that when it is not only one learner the appropriate method of intervention may differ. Although the present findings show potential for improving the adaptivity and effectiveness of intelligent tutoring systems and other learning environments, there are still many questions about which intervention to deploy when in the learning session and to whom.

Limitations and Future Directions

Although the present dissertation attempted to completely investigate interventions to regulate confusion during learning, there are limitations to each of the four studies that were conducted. Overall, each of the studies was only one investigation into that particular aspect of confusion regulation interventions. Replications are needed for each study to determine the reliability of these findings and also to determine the conditions under which each of these findings occurs.

For the learner preferences study, there was one important limitation. The limitation was that each emotion was not defined for the learners. Thus, it may have been

the case that what one learner viewed as confusion was what another learner viewed as frustration. This is particularly important given that persistent confusion that cannot be resolved can transition into frustration (D'Mello & Graesser, 2012b). Thus, the less intense experiences of confusion might be easily distinguished from frustration, whereas more intense experiences of confusion may be highly similar to frustration. In addition to defining emotions for learners, it would also be beneficial to assess how learners generally perceive confusion (e.g., learning opportunity, indicates lack of skills, etc.).

For the expert tutor study, there was also one important limitation. The limitation was that the tutors were not consulted to determine what strategies they were adopting and what they were actually responding to during the tutoring session (e.g., cognitive state only, affective state only, cognitive + affective state). In future studies, it would be helpful to have tutors go through a session and prompt them to indicate what their thought process was at critical points in the tutoring session. This would enable future studies to determine if there were larger strategies being employed by the tutors, if tutors were aware of and responding to learner emotions, and if there were other learner characteristics that tutors were taking into consideration when providing tutorial instruction.

For the pedagogical and motivational intervention studies, there were four important limitations. First, critics might object to the confusion induction manipulation on the grounds that learners were provided with intentionally misleading information and contradictions and this is not in their best interest. This concern and similar reactions to the manipulation are acknowledged, but the present dissertation takes the position that these are less of a concern in the present research for the following reasons: (a) any

misleading information presented was corrected at the end of the experiment, (b) all research protocols were approved by the appropriate IRB board, (c) learners were consenting research participants instead of actual students, and (d) learners were fully debriefed at the end of the experiment.

The second limitation addresses the ability to apply the present techniques to a broader range of domains. Both Studies 3 and 4 were conducted within the domain of research methods and utilized discussions of research case studies to help learners reach a better understanding of research methods concepts. It must be plausible to have disagreements during a discussion in order to utilize this method of confusion induction. However, the confusion regulation interventions may be applicable to a greater number of domains. This may be the case due to the fact that constructing an argument to convince a hypothetical new student was also found to be effective. Therefore, it is not necessary for the agent(s) in the learning environment to pose disagreements in order to make use of the confusion regulation interventions.

The third and fourth limitations consider the nature of the interventions to regulate confusion. The interventions provided in the current experiments could be improved in a number of ways. For the pedagogical interventions, for example, it may be the case that a more interactive intervention (e.g., scaffolding to address specific misconceptions or errors) may have been more effective. As mentioned previously, the motivational interventions could be improved by devoting more time to retraining learner attributions about confusion as well as assessing current learner attributions to tailor the intervention more to the specific learner. Finally, there are other types of pedagogical and motivational interventions that could have been used as well as other methods of

intervention that could, and should, be explored. For example, the interventions that learners rated during the learner preferences study could be investigated within the confusion induction learning environment.

There are also other components of the learning interaction that need to be further investigated in addition to other types of confusion regulation interventions. One important area is to investigate learner characteristics that impact the effectiveness of confusion induction methods and confusion regulation interventions. For example, in a previous study that used false feedback to induce confusion, it was found that learners with high prior knowledge and high cognitive drive (e.g., prefer difficult material, enjoy challenging material, prefer complex explanations, etc.) were more likely to be successfully confused by false feedback, spent more time in the confusion regulation task (reading an explanatory text), and performed better on transfer tasks when they were in a confusion regulation condition (Lehman et al., 2013). Thus, it is important to explore the relationship between learner characteristics and the findings from the learner preferences study and the two confusion regulation intervention studies. It could be, for example, that the Read Only condition performed surprisingly well because there were many learners in that condition that have higher self-motivation to work towards resolving their confusion. It is also important to identify those learners that do not need an intervention. An effective learning must be able to determine both when it is necessary to intervene and when it is best to let the learner work through confusion on their own.

Conclusion

Overall the present dissertation has found evidence that confusion regulation interventions can be helpful for learning. However, it does not seem to be that a simple

one-size-fits-all approach can be adopted. Adaptive learning environments will need to determine which confusion regulation intervention will be most effective based on the learner's characteristics and their current performance in the learning session. Another aspect of confusion regulation that was not addressed in the present dissertation but is important is approaches that can be used in the classroom. Tutors and adaptive learning environments can provide one-to-one instruction that can be adapted to a particular learner, but what does a teacher with thirty students in the classroom do? It is important for future research to determine which strategies work for which learners, but also which strategies work in which contexts. In order to promote deeper learning, teachers, tutors, and adaptive learning environments will need to adapt to both learner cognitive and affective states.

References

Anderson, C. A. (1983). Motivational and performance deficits in interpersonal settings: The effects of attributional style. *Journal of Personality and Social Psychology*, *45*, 1136-1147.

Andrews, G. R., & Debys, R. L. (1978). Persistence and the causal perception of failure: Modifying cognitive attributions. *Journal of Educational Psychology*, *70*, 154-166.

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.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (Release 2) [CD-ROM]. University of Pennsylvania, Linguistic Data Consortium, Philadelphia, PA.

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

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

Batson, C. D., Turk, C. L., Shaw, L. L., & Klein, T. R. (1995). Information function of empathic emotion - Learning that we value the others welfare. *Journal of Personality and Social Psychology*, *68*, 300-313.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*, 539-546.

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

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

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, *4*, 215-222.

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

Cade, W., Lehman, B., & Olney, A. (2010). An exploration of off topic conversation. In J. Burstein, M. Harper, & G. Penn (Eds.), *NAACL-HLT 2010 Conference*. Los Angeles, CA.

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

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

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

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

Clifford, M. M. (1984). Thoughts on a theory of constructive failure. *Educational Psychologist*, *19*, 108-120.

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

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

D'Esposito, M., Postle, B. R., & Rypma, B. (2000). Prefrontal cortical contributions to working memory: Evidence from event-related fMRI studies. *Experimental Brain Research*, *133*, 3-11.

D'Mello, S. K. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology, *Journal of Educational Psychology, 105*, 1082-1099.

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

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

D'Mello, S. K., & Graesser, A. C. (2012a). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems*, *2*(4), 23: 1-38.

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

D'Mello, S. K., Lehman, B. A., & Graesser, A. C. (2011). A motivationally supportive affect-sensitive AutoTutor. In R. A. Calvo & S. K. D'Mello (Eds.), *New perspectives on affect and learning technologies* (pp. 113-126). New York: Springer.

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

D'Mello, S., Lehman, B., & Person, N. (2010). Expert tutors feedback is immediate, direct, and discriminating. *Proceedings of the 23rd Florida Artificial Intelligence Research Society Conferences (FLAIRS-23)* (pp. 595-604). AAAI Press.

Damasio, A. R. (1994). *Descartes' error*. New York, NY: Putnam.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*, 1087-1101.

Dunlosky, J., & Lipko, A. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*, 228-232.

Dweck, C. (1999). *Self theories: Their role in motivation, personality and development*. Philadelphia, PA: Taylor & Francis/Psychology Press.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion in the human face: guide-lines for research and an integration of findings*. New York, NY: Pergamon Press.

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

Fletcher, G. J. O., Fernandez, G., Peterson, D., & Reeder, G. D. (1986). Attributional complexity: An individual differences measure. *Journal of Personality & Social Psychology*, *51*, 875-884.

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

Forbes-Riley, K., & Litman, D. (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. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, *53*, 1115-1136.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 702-718.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, *10*, 597-602.

Graesser, A.C., & D'Mello, S. (2012). Emotions during the learning of difficult material. In. B. Ross (Eds.), *The Psychology of Learning and Motivation*, vol. 57 (183-225). Elsevier.

Graesser, A. C., D'Mello, S. K., Craig, S. D., Witherspoon, A., Sullins, J., McDaniel, B., & Gholson, B. (2008). The relationship between affect states and dialogue patterns during interactions with AutoTutor. *Journal of Interactive Learning Research*, *19*, 293–312.

Graesser, A. C., D'Mello, S. K., & Person, N. K. (2009). Meta-knowledge in tutoring. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 361-382). New York, NY: Routledge.

Graesser, A., Lu, S. L., Jackson, G., Mitchell, H., Ventura, M., Olney, A., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*, *36*, 180-193.

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

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). (2009). Handbook of metacognition in education. New York, NY: Routledge.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*, 10.

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

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

Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: John Wiley & Sons.

Izard, C. E. (1993). Four systems for emotion activation: Cognitive and non-cognitive processes. *Psychological Review*, *100*, 68-90.

Lehman, B., Cade, W., & Olney, A. (2010). Off topic conversation in expert tutoring: Waste of time or learning opportunity? In S.J.d. Baker, A. Merceron, & P. Pavlik (Eds.), *Proceedings of 3rd International Conference on Educational Data Mining*.

Lehman, B., D'Mello, S., & Graesser, A. (in preparation). False feedback can improve learning when you're productively confused.

Lehman, B., D'Mello, S., & Graesser, A. (2013). Who benefits from confusion during learning? An individual differences cluster analysis. In K. Yacef, C. Lane, J. Mostow, & P. Pavlik (Eds.*), Proceedings of 16th International Conference on Artificial Intelligence in Education (AIED2013)* (pp. 51-60). Berlin Heidelberg: Springer-Verlag.

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

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

Lepper, M., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 135-158). Orlando, FL: Academic Press.

Litman, D., & Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, *48*, 559–590.

Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117-144). Mahwah, NJ: Erlbaum.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2004). Emotional intelligence: Theory, findings, and implications. *Psychological Inquiry*, *15*, 197-215.

McQuiggan, S., Mott, B., & Lester, J. (2008). Modeling self-efficacy in intelligent tutoring systems: an inductive approach. *User Modeling and User-Adapted Interaction*, *18*, 81-123.

McQuiggan, S., Robison, J., & Lester, J. (2010). Affective transitions in narrative-centered learning environments. *Educational Technology & Society*, *13*, 40-53.

Medway, F. J., & Venino, G. R. (1982). The effects of effort-feedback and performance patterns on children's attributions and task persistence. *Contemporary Educational Psychology*, *7*, 26-34.

Ochsner, K. N., & Gross, J. J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences*, *9*, 242-249.

Pekrun, R. (2010). Academic emotions. In T. Urdan (Ed.), *APA educational psychology handbook* (Vol. 2). Washington, DC: American Psychological Association.

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

Perry, R. P., Stupinsky, R. H., Hall, N. C., Chipperfield, J. G., & Weiner, B. (2010). Bad starts and better finishes: Attributional retraining and initial performance in competitive achievement settings. *Journal of Social and Clinical Psychology*, *29*, 668-700.

Person, N., Lehman, B., & Ozbun, R. (2007). Pedagogical and motivational dialogue moves used by expert tutors. *Paper presented at the 17th Annual Meeting of the Society for Text and Discourse*. Glasgow, Scotland.

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

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

Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulation learning components of classroom academic performance. *Journal of Educational Psychology*, *82*, 33-40.

Reisenzein, R. (1983). The Schachter theory of emotion: Two decades later. *Psychological Bulletin*, *94*, 239-264.

Robison, J., McQuiggan, S., & Lester, J. (2009). Evaluating the consequences of affective feedback in intelligent tutoring systems. In J. Cohn, A. Nijholt, & M. Pantic (Eds.), *Proceedings of the International Conference on Affective Computing & Intelligent Interaction* (pp. 37-42). Amsterdam: IEEE.

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

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

Rogoff, B., & Gardner, W. (1984). Adult guidance of cognitive development. In B. Rogoff & J. Lave (Eds.), *Everyday cognition: Its development in social context* (pp. 95-116). Cambridge, MA: Harvard University Press.

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

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, *92*, 447-472.

Schachter, S., & Singer, J. E. (1962). Cognitive, social and psychological determinants of emotional states. *Psychological Review*, *88*, 408-437.

Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition & Emotion*, *7*, 325-355.

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.

Strain, A., D'Mello, S. K., & Gross, M. (2012). How Do Learners Regulate their Emotions? In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.) *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 618-619). Berlin Heidelberg: Springer-Verlag.

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

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*, 209-249.

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

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York, NY: Springer- Verlag.

Weiner, B. (2010). The development of an attribution-based theory of motivation: A history of ideas. *Educational Psychologist*, *45*, 28-36.

Wilson, T. D., & Linville, P. W. (1982). Improving the academic performance of college freshmen: Attribution therapy revisited. *Journal of Personality and Social Psychology*, *42*, 367-376.

Wilson, T. D., & Linville, P. W. (1985). Improving the performance of college freshmen with attributional techniques. *Journal of Personality and Social Psychology*, *49*, 287-293.

Zoeller, C. J., Mahoney, G., & Weiner, B. (1983). Effects of attribution training on the assembly task performance of mentally retarded adults. *American Journal of Mental Deficiency*, *88*, 109-112.

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, *39*, 35-62.

Appendix A – Survey Research Questions (Study 1)

INSTRUCTIONS:
You will be asked several questions about different emotional experiences (boredom, confusion, frustration) during learning. Learning refers to any experience in a classroom, working alone, or working with a tutor in which you are attempting to learn some material. There are not right or wrong answers. Please answer honestly and to the best of your ability.

1. In general, what type of assistance or learning intervention do you think would help you or other learners to overcome CONFUSION during learning?

2. In general, what type of assistance or learning intervention do you think would help you or other learners to overcome BOREDOM during learning?

3. In general, what type of assistance or learning intervention do you think would help you or other learners to overcome FRUSTRATION during learning?

Rate each question on the following scale:

1	2	3	4	5	6
Not at All Helpful	Not Helpful	Somewhat Not Helpful	Somewhat Helpful	Helpful	Very Helpful

1. When you are CONFUSED during learning, how helpful would you find each of the following for overcoming your CONFUSION?
    a. Additional information about the concept being learned
    b. Encouragement to persist with the task
    c. Presentation of a new (but related) task to solve
    d. Feedback about the quality of your progress/responses (i.e., correct vs. incorrect)
    e. Statement of the correct answer
    f. A short break to do an unrelated task

2. When you are BORED during learning, how helpful would you find each of the following for overcoming your BOREDOM?
    a. Additional information about the concept being learned
    b. Encouragement to persist with the task
    c. Presentation of a new (but related) task to solve
    d. Feedback about the quality of your progress/responses (i.e., correct vs. incorrect)
    e. Statement of the correct answer
    f. A short break to do an unrelated task

3. When you are FRUSTRATED during learning, how helpful would you find each of the following for overcoming your FRUSTRATION?
    a. Additional information about the concept being learned
    b. Encouragement to persist with the task
    c. Presentation of a new (but related) task to solve
    d. Feedback about the quality of your progress/responses (i.e., correct vs. incorrect)
    e. Statement of the correct answer
    f. A short break to do an unrelated task

Appendix B – Informed Consent Form (Study 1)

AGREEMENT TO PARTICIPATE

In this study, you will be asked to complete several surveys about your emotional experiences during learning.

You will also be asked to complete several surveys about your learning experiences in general and a demographics questionnaire.

The duration of the study is approximately 60 minutes. You will receive 1 credit for taking part in the study.

Your participation is voluntary, and you are free to withdraw from the research at any time. If you withdraw from the study, you will receive credit for the portion of the study that you completed.

Participation in this study should not pose any risk.

To participate in this study and receive credit, you must sign this form.

By electronically signing below, you agree to participate in the proposed study and confirm that you have read this agreement. If you have any further questions regarding your participation or any other study-related questions, please contact Blair Lehman (balehman@memphis.edu). If you have any questions regarding your rights as a subject in this study, you may contact the chair of the Institutional Review Board for the Protection of Human Subjects at 901-678-2533.

Electronic Signature (Name):

Date:

Appendix C – Data Release Agreement Form (Study 1)

DATA RELEASE AGREEMENT

I agree to let my data (responses to survey questions) be used for presentation at conferences, in journal publication, and book chapters. This information will only be used as examples of data output.

I agree to let my data be used in future studies. This would involve a new set of participants viewing parts of my data. These participants will sign a confidentiality agreement to viewing my data.

I understand that my personal data will never be associated with my personal identification information. I understand that agreement to this usage is completely voluntary and I am able at any point to refuse my data be used in this way.

You do not have to sign this form to participate in this study and receive credit.

Electronic Signature (Name):

Date:

Appendix D – Debriefing Form (Study 1)

DEBRIEFING

The purpose of this study was to learn more about how people interpret emotional experiences during learning and how these experiences relate to their more general emotional experiences. We focused on boredom, confusion, and frustration because these are emotions that frequently occur during learning and often require some type of instructional or motivational intervention to help learners succeed.

In future research, we hope to develop and test interventions that are adaptive to both learners' cognitive and affective states. The present survey will help us to determine how these different emotional experiences make learners feel about themselves. This information will then allow us to develop interventions that are more appropriate for learners and hopefully more effective.

Your data will only be anonymously shared with credible researchers and all information collected in this study will be kept confidential within the limits of law.

We thank you for your participation in this study. Your contribution was instrumental in advancing our knowledge of how emotions impact deep learning gains. The results of this study will be used to engineer a computer tutor that promotes, tracks, and helps regulate emotions.

For more information on this project or if you have any questions and concerns, please contact Blair Lehman (balehman@memphis.edu).

Appendix E – Academic Grit Scale

There are no right or wrong answers. We are interested in your own perceptions. Please answer each question as honestly and accurately as you can, but don't spend too much time thinking about each answer. Select an answer option based on how much each statement applies to you (1 = Not At All Like Me, 5 = Very Much Like Me).

1. I often set a goal but later choose to pursue a different one.
2. New ideas and new projects sometimes distract me from previous ones.
3. I become interested in new pursuits every few months.
4. My interests change from year to year.
5. I have been obsessed with a certain idea or project for a short time but later lost interest.
6. I have difficulty maintaining my focus on projects that take more than a few months to complete.
7. I have achieved a goal that took years of work.
8. I have overcome setbacks to conquer an important challenge.
9. I finish whatever I begin.
10. Setbacks don't discourage me.
11. I am a hard worker.
12. I am diligent.

Appendix F – School Failure Tolerance Scale

There are no right or wrong answers. We are interested in your own perceptions. Please read over each statement carefully and think about how much it does or does not apply to you. Select an option that best applies to you. (1 = Extremely Disagree, 3 = Somewhat Disagree, 4 = Somewhat Agree, 6 = Extremely Agree)

1. I feel terrible when I make a mistake in school.
2. If I do poorly in my school work, I try not to let anyone know.
3. A low grade in my school work makes me feel very sad.
4. When I start something new in school, the first think I think about is that I might fail.
5. I worry a lot about making errors in my school work.
6. I feel like hiding whenever I get a bad grade in school.
7. If I make lots of mistakes in school, I feel very moody or angry.
8. I don't like to study with classmates because they may think I am dumb if I don't know something.
9. When I fail at something in school, I don't like to eat, or play, or talk, or do anything.
10. I get very discouraged if I make errors on a task I am trying to learn.
11. I really dislike school work on which I make mistakes.
12. If I give a wrong answer to a teacher's question, I feel terrible.
13. I like to do school work that is difficult for me.
14. I would rather work problems I can do in a hurry that those that take much time and thought.
15. I would do almost anything to get out of working difficult problems in school.
16. I like to try difficult assignments even if I get some wrong.
17. School work that really makes me think is fun.
18. School work that is difficult is more fun than work that is very easy.
19. I would rather study a difficult course than a very easy one.
20. If I could choose my math problems, I would pick hard ones rather than very easy ones.
21. It is fun to try to answer questions that are difficult or challenging.
22. The easier school work is for me, the more I like it.
23. I like to study with classmates that enjoy working on difficult lessons.
24. I would rather make mistakes on a difficult task than get a perfect score on an easy but boring task.
25. I like to ask questions in school because I learn by asking questions.
26. If I can't succeed at a new school task, I give up quickly.
27. When I make mistakes in my school work, I just keep trying and trying.
28. I don't like to set goals for my school work, because I might not reach them and then I feel bad.
29. If a school task is difficult, I try to get by without doing it.
30. If I do not understand something, I ask the teacher to explain it.
31. I would rather guess at something and get it wrong than ask a question that may sound silly.

32. I almost always learn a lot from the mistakes I make in my school work.
33. If I get a low grade in my school work, I study my errors and rework the problems I get wrong.
34. I usually study and correct the errors I make on school work, even if I don't have to.
35. I don't like to set goals for my school work. I just do the work and forget about it.
36. If I get a low score, I usually make up my mind to buckle down and study hard.

Appendix G – Motivated Strategies for Learning Questionnaire

Read every question and select the option that suits you best. There are no correct or incorrect answers. Choose one option per question. Work quickly.

Indicate the degree to which each statement is true of you. (1 = Not at All True of Me, 3 = Somewhat True of Me, 4 = Neither True nor Not True of Me, 5 = Somewhat True of Me, 7 = Very True of Me)

1. Compared with other students in my classes I expect to do well.
2. I'm certain I can understand the ideas taught in my courses.
3. I expect to do very well in my classes.
4. Compared with others in my classes, I think I'm a good student.
5. I am sure I can do an excellent job on the problems and tasks assigned for my classes.
6. I think I will receive good grades in my classes.
7. My study skills are excellent compared with others in my classes.
8. Compared with other students in my classes I think I know a great deal about the subjects.
9. I know that I will be able to learn the material for my classes.
10. I prefer class work that is challenging so I can learn new things.
11. It is important for me to learn what is being taught in my classes.
12. I like what I am learning in my classes.
13. I think I will be able to use what I learn in my classes in other classes.
14. I often choose paper topics I will learn something from even if they require more work.
15. Even when I do poorly on a test I try to learn from my mistakes.
16. I think that what I am learning in my classes is useful for me to know.
17. I think that what we are learning in my classes is interesting.
18. Understanding the subject taught in my classes is important to me.
19. I am so nervous during a test that I cannot remember facts I have learned.
20. I have an uneasy, upset feeling when I take a test.
21. I worry a great deal about tests.
22. When I take a test I think about how poorly I am doing.
23. When I study for a test, I try to put together the information from class and from the book.
24. When I do homework, I try to remember what the teacher said in class so I can answer the questions correctly.
25. It is hard for me to decide what the main ideas are in what I read.
26. When I study I put important ideas into my own words.
27. I always try to understand what the teacher is saying even if it doesn't make sense.
28. When I study for a test I try to remember as many facts as I can.
29. When studying, I copy my notes over to help me remember material.
30. When I study for a test I practice saying the important facts over and over to myself.

31. I use what I have learned from old homework assignments and the textbook to do new assignments.
32. When I am studying a topic, I try to make everything fit together.
33. When I read material for my classes, I say the words over and over to myself to help me remember.
34. I outline the chapters in my book to help me study.
35. When reading, I try to connect the things I am reading about with what I already know.
36. I ask myself questions to make sure I know the material I have been studying.
37. When work is hard I either give up or study only the easy parts.
38. I work on practice exercises and answer end of chapter questions even when I don't have to.
39. Even when study materials are dull and uninteresting, I keep working until I finish.
40. Before I begin studying I think about the things I will need to do to learn.
41. I often find that I have been reading for class but don't know what it is all about.
42. I find that when the teacher is talking I think of other things and don't really listen to what is being said.
43. When I'm reading I stop once in a while and go over what I have read.
44. I work hard to get a good grade even when I don't like a class.

Appendix H – Attributional Complexity Scale

There are no right or wrong answers. We are interested in your own perceptions. Please answer each question as honestly and accurately as you can, but don't spend too much time thinking about each answer. Select an answer option based on how much you agree or disagree that statement applies to you. (-3 = Strongly Disagree, 0 = Neither Agree nor Disagree, 3 = Strongly Agree)

1. I don't usually bother to analyze and explain people's behavior.
2. Once I have figured out a single cause for a person's behavior I don't usually go any further.
3. I believe it is important to analyze and understand our own thinking processes.
4. I think a lot about the influence that I have on other people's behavior.
5. I have found that relationships between a person's attitudes, beliefs, and character traits are usually simple and straightforward.
6. If I see people behaving in a really strange or unusual manner I usually put it down to the fact that they are strange or unusual people and don't bother to explain it any further.
7. I have though a lot about the family background and personal history of people who are close to me, in order to understand why they are the sort of people they are.
8. I don't enjoy getting into discussions where the causes for people's behavior are being talked over.
9. I have found that the causes for people's behavior are usually complex rather than simple.
10. I am very interested in understanding how my own thinking works when I make judgments about people or attach causes to their behavior.
11. I think very little about the different ways that people influence each other.
12. To understand a person's personality/behavior I have found it is important to know how that person's attitudes, beliefs, and character traits fit together.
13. When I try to explain other people's behavior I concentrate on the person and don't worry too much about all the existing external factors that might be affecting them.
14. I have often found that the basic cause for a person's behavior is located far back in time.
15. I really enjoy analyzing the reasons or causes for people's behavior.
16. I usually find that complicated explanations for people's behavior are confusing rather than helpful.
17. I give little thought to how my thinking works in the process of understanding or explaining people's behavior.
18. I think very little about the influence that other people have on my behavior.
19. I have thought a lot about the way that different parts of my personality influence other parts (e.g., beliefs affecting attitudes or attitudes affecting character traits).
20. I think a lot about the influence that society has on other people.
21. When I analyze a person's behavior I often find the causes from a chain that foes back in time, sometimes for years.

22. I am not really curious about human behavior.
23. I prefer simple rather than complex explanations for people's behavior.
24. When the reasons I give for my own behavior are different from someone else's, this often makes me think about the thinking processes that lead to my explanations.
25. I belief that to understand a person you need to understand the people who that person has close contact with.
26. I tend to take people's behavior at face value and not worry about the inner causes for their behavior (e.g. attitudes, beliefs, etc.).
27. I think a lot about the influence that society has on my behavior and personality.
28. I have thought very little about my own family background and personal history in order to understand why I am the sort of person I am.

Appendix I – Demographics Questionnaire

Please answer the following questions to the best of your ability.

1. What is your current age in years?
2. What is your gender?
    a. Male
    b. Female
3. Which ethnicity best describes you?
    a. African-American/Black
    b. Caucasian/White
    c. Hispanic, Latino, or Mexican origin
    d. Asian
    e. American Indian/Alaskan Native
    f. Native Hawaiian or other Pacific Islander
    g. Other
4. What year are you currently in?
    a. First Year
    b. Sophomore
    c. Junior
    d. Senior
5. What was your ACT or SAT score?

Appendix J – Informed Consent Form (Studies 3 & 4)

In this study, you will be asked to use a computer-based learning environment to learn important critical thinking skills.

You will also complete a pretest, posttest, and a few questionnaires throughout this study. We will record a video of your face, your speech, your computer screen, mouse movements, and responses to the tutor. This data will only be used by the research team and not shared with anyone without your expressed written consent.

The duration of the study is approximately 120 minutes. You will receive 2 hours of research credit for taking part in the study.

Your participation is voluntary, and you are free to withdraw from the research at any time. If you withdraw from the study, you will receive credit for the time you remained in the study.

Participation in this study should not pose any risk.

By signing below, you agree to participate in the proposed study and confirm that you have read and received a copy of this agreement. If you have any further questions regarding your participation or any other study-related questions, please contact Blair Lehman (balehman@memphis.edu). If you have any questions regarding your rights as a subject in this study, you may contact the chair of the Institutional Review Board for the Protection of Human Subjects at 678-2533.

_____

Your Signature

_____

Your Printed Name

_____

Your Email Address

_____

Today's Date

Appendix K – Data Release Agreement Form (Studies 3 & 4)

I agree to let my data (videos of face, audio, video of computer screen, mouse movements, emotion measures, and knowledge measures) be used for presentation at conferences, in journal publication, and book chapters. This information will only be used as examples of data output.

I agree to let my data be used in future studies. This would involve a new set of participants viewing parts of my data. These participants will sign a confidentiality agreement prior to viewing my data.

I understand that my personal data will never be associated with my personal identification information or test scores. I understand that agreement to this usage is completely voluntary and I am able at any point to refuse my data be used in this way.

_____

Your Signature

_____

Your Printed Name

_____

Today's Date

_____

Experimenter Printed Name

_____

Experimenter Signature

Appendix L – Debriefing Form (Study 3)

The purpose of this study was to experiment with different tutorial interventions to help you obtain a deep understanding of topics in critical thinking. We focused on critical thinking because it is widely acknowledged that the level of science understanding in the United States is unacceptably low, yet the advancement of scientific knowledge depends on the application of skills needed for scientific inquiry. Our research aspires to fill this gap by developing technological interventions to fortify citizens and aspiring scientists with the skills needed for critical thinking, model-based reasoning, and problem solving in science.

Decades of previous research, have revealed that students rarely acquire a deep understanding of difficult conceptual information from reading the textbook or traditional classroom instruction. Hence, we explored a different strategy to help you learn. In particular, we tried to confuse you so that you would stop and think. This is because our previous research indicates that a degree of confusion is essential for learning, particularly at deeper levels of comprehension. So the tutor tried to confuse you by providing misleading information and contradicting you in conjunction with the student agent. Note that all misleading information was eventually corrected over the course of the tutoring session. In addition to confusing you, the agent also attempted to regulate this confusion as well through different types of interventions (i.e., construct a convincing argument, read an explanatory text, construct a convincing argument with the aid of an explanatory text, attempt to construct a convincing argument and then read an explanatory text). You received one of these types of interventions to test whether this could increase your level of learning.

You were asked to take a knowledge test before interacting with the tutor and another knowledge test after interacting with the tutor. We will use the difference between your pre and post test scores to calculate your learning gains. We will test to see if you learned more when you were confused compared to when you were not confused.

We recorded a video of your face, a video of your computer screen, your responses to the tutor's questions, and your mouse movements. This data will be used to explore connections between these various channels and your levels of confusion.

Your data will only be anonymously shared with credible researchers and all information collected in this study will be kept confidential within the limits of law.

We thank you for your participation in this study. Your contribution was instrumental in advancing our knowledge of how confusion impacts deep learning gains. The results of this study will be used to engineer a computer tutor that promotes, tracks, and helps regulate confusion during learning.

For more information on this project or if you have any questions or concerns, please contact Blair Lehman (balehman@memphis.edu).

Appendix M – Debriefing Form (Study 4)

The purpose of this study was to experiment with different tutorial interventions to help you obtain a deep understanding of topics in critical thinking. We focused on critical thinking because it is widely acknowledged that the level of science understanding in the United States is unacceptably low, yet the advancement of scientific knowledge depends on the application of skills needed for scientific inquiry. Our research aspires to fill this gap by developing technological interventions to fortify citizens and aspiring scientists with the skills needed for critical thinking, model-based reasoning, and problem solving.

Decades of previous research, have revealed that students rarely acquire a deep understanding of difficult conceptual information from reading the textbook or traditional classroom instruction. Hence, we explored a different strategy to help you learn. In particular, we tried to confuse you so that you would stop and think. This is because our previous research indicates that a degree of confusion is essential for learning, particularly at deeper levels of comprehension. So the tutor tried to confuse you by providing misleading information and contradicting you in conjunction with the student agent. Note that all misleading information was eventually corrected over the course of the experiment. In addition to confusing you, the agent also attempted to regulate this confusion as well through different types of motivating interventions (i.e., general motivational statement, attribute confusion to difficulty of the material with a motivational statement, attribute confusion to tutor explanations with a motivational statement, or reframe the confusion experience with a motivational statement). You received one of these interventions to test whether this could increase your level of learning.

You were asked to take a knowledge test before interacting with the tutor and another knowledge test after interacting with the tutor. We will use the difference between your pre and post test scores to calculate your learning gains. We will test to see if you learned more when you were confused compared to when you were not confused.

We recorded a video of your face, a video of your computer screen, your responses to the tutor's questions, and your mouse movements. This data will be used to explore connections between these various channels and your levels of confusion.

Your data will only be anonymously shared with credible researchers and all information collected in this study will be kept confidential within the limits of law.

We thank you for your participation in this study. Your contribution was instrumental in advancing our knowledge of how confusion impacts deep learning gains. The results of this study will be used to engineer a computer tutor that promotes, tracks, and helps regulate confusion during learning.

For more information on this project or if you have any questions or concerns, please contact Blair Lehman (balehman@memphis.edu).

Appendix N – Flaw-Identification Task Pretest (Studies 3 & 4)

Please read over the experiment carefully and take your time responding to each choice. Click on those elements that you feel are a problem and leave blank those elements that you think are not a problem. In each experiment there may be no problem, one problem, or more than one problem.

___ Construct Validity
___ Control Group
___ Correlational Study
___ Experimenter Bias
___ Generalizability
___ Measurement Sensitivity
___ Random Assignment
___ Replication
___ No Problem

Many people claim that punishing children is bad parenting and can cause children to behave worse in the future. A group of researchers tested this claim. They had parents and their children come into a laboratory room with two-way mirrors. Parents were put into either "punishment" (say whatever is necessary to reprimand the child) or "no punishment" condition (say nothing). The researchers then created a situation in which the child behaved badly. One researcher would call a parent over to the door for a quick conversation, while a second researcher would give the child a permanent marker. The child then draw on the wall (time 1). Parents and children came back a week later and were put in the same situation (time 2), but the parents were allowed to chose whether or not to punish their child. There were two measures: (1) parents' behavior (at both times) and (2) whether or not the child drew on the wall. The researchers found that there was no difference in the children's behavior at time 2 and that parents' punishment was not severe at either time. So the researchers concluded that punishment is not detrimental to children.

Half of the students in a learning disability class received the test-taking pro treatment and half didn't. The treatment teaches students about breaking down test questions into smaller, more manageable pieces. The researchers that developed the treatment worked with these students. They wanted to see if the treatment really worked, so they ran a study to evaluate test-taking skills. All of the students took a test and then reported how many questions they thought they answered correctly. The researchers wanted to make sure they had the same amount of data for each student, so the researchers stood nearby while the students took the test to answer any questions that the students had. The researchers looked at the data and found that the test-taking pro treatment did work.

There's this new anti-depression pill called "Lighten Up" and people had a 20% reduction in their depression level in the first two weeks of taking it. None of the participants had therapy sessions or any other type of alternative treatment for depression, they just went on with their usual routines. The results were compared to another group of people who didn't take the pill. The group that didn't take Lighten Up had no reduction in their depression level. So Lighten Up works a lot better than just dealing with depression yourself.

A past study showed that people who drive sports cars are more reckless in their behavior than people who drive other types of cars. A group of researchers were skeptical about this finding, so they ran their own study to investigate this. So here's the study that they ran. They took people who drove sports cars and other types of cars and had them fill out a survey. The survey asked questions about drug use, since drug use itself is a reckless behavior and can lead to other types of reckless behaviors, and what people do while on drugs. All of the participants filled out this survey and did not know what the researchers were investigating. The surveys were scored for amount and intensity of reckless behavior. The researchers found that there were not significant differences in reckless behavior between sports car drivers and people who drive other types of cars. So the previous study must have been incorrect.

One school year a principal decided to figure out if student manuals actually helped students to follow all of the school rules and avoid detentions and suspensions. The new students to high school could choose to take biology the first or second semester their freshman year. So the principal had the biology teacher go over the importance of the student manual during the fall semester, but only said that the manual was optional in the spring semester. The principal found that there were no differences in the amount of behavioral problems between the two groups of students. So to save money she has stopped printing the student manual.

A high school teacher wanted to see if just telling her students that they were drinking caffeinated coffee, when they actually got decaf, would give them an energy boost like actually drinking caffeinated coffee. And that's what happened! Here's how she did her experiment, 50 students drank the decaf coffee and 50 drank the caffeinated coffee. She found students felt the same, but she wasn't sure if she could trust her findings, so she ran the experiment again with the same students the next semester. To make sure it really worked and her students wouldn't catch on, she ran the experiment again the next semester. She made sure to randomly assign the same 100 students to the two groups (caffeinated or decaf coffee) each time. The decaf worked just as well again! So she thinks you should just change the labels on the coffee beans.

Appendix O – Near Transfer Task Posttest (Studies 3 & 4)

Please read over the experiment carefully and take your time responding to each choice. Click on those elements that you feel are a problem and leave blank those elements that you think are not a problem. In each experiment there may be no problem, one problem, or more than one problem.

___ Construct Validity
___ Control Group
___ Correlational Study
___ Experimenter Bias
___ Generalizability
___ Measurement Sensitivity
___ Random Assignment
___ Replication
___ No Problem

One year when hiring new employees a boss decided to test if the training manuals were really any help. So he told the employees who began work in April to not use the training manuals and he told the employees who began work in June to use the training manuals. After each group of employees had been working for six months, he evaluated their overall job performance. He found that there weren't any differences between the employees who were told to use the training manual and the employees who were told not to use the training manual. So he concluded that it's a waste of company money to make these training manuals and decided that everyone can just get on-the-job-training and figure things out for themselves.

Does your child get really scared during thunder and lightning storms? Well now there's this great new meditation exercise that will help your child to feel calm and relaxed or even go to sleep during these storms! Researchers did a study on the new meditation exercise, "Fear-B-Gone". These researchers found that after children performed this 10-minute exercise they immediately reported reduced levels of fear during thunder and lightning storms. The article said that all of the children in the study stopped using any other strategies for coping with this fear when they started doing the "Fear-B-Gone" meditation. They didn't do anything besides the meditation. There was another group of children who didn't do the "Fear-B-Gone" meditation and continued with their normal routines for dealing with their fear. The group that didn't perform the "Fear-B-Gone" meditation had no reduction in their fear during thunder and lightning storms. So "Fear-B-Gone" is a quick, new miracle solution for this fear!

A kindergarten teacher realized that the biggest problem with her students was separation anxiety. So she enrolled half her students in a program specializing in separation anxiety. Only half the students were enrolled so she could see if the program actually helped. The researchers in charge of the program were excited to have a teacher's full cooperation, so they ran an experiment to see how effective the program really was. The researchers spoke to students about how they felt being away from home and to rate their current feelings. The researchers often had to prompt students to get more detailed responses about how they were feeling. Students who had participated in the program behaved significantly better in class and dealt with their feelings of separation anxiety much better than the other students.

Recently a company has started selling energy cookies to give people an energy boost in the form of a snack. A boss recently noticed that a lot of his employees were eating these energy cookies, but these energy cookies are pretty expensive. So the boss gave half of his employees the real energy cookies and the other half he gave regular cookies that looked identical to the energy cookies, this way all of the employees thought they were getting energy cookies. All of his employees felt an energy boost, even the ones who just ate regular cookies! The boss wanted to make sure that he didn't just pick a day where everyone felt energized because of some other reason, like doing well at work. So he ran the study again using all of his employees. He flipped a coin to decide which employees got the energy cookies and which ones got the regular cookies in both studies. Then he ran the exact same study again. The second study showed the exact same results as the first study. So the boss decided that he'll just bring in regular cookies every day but tell his employees that they're energy cookies. This way he'll have more efficient employees and save money!

About a decade ago, there was a study that looked at random acts of kindness. This study found that children under the age of 12 were more likely to do random acts of kindness than children between 13 and 18 years old. A researcher had doubts that this was really true, so she designed a study to test this out. To test this, she videotaped children at recess and other breaks during the school day (e.g., lunch, passing periods, etc.). She then had another researcher code the videos for random acts of kindness. The researcher who did the coding was told to count the number of times an individual child said please and thank you to another child and teachers. After looking at videos from many schools in different parts of the US, she found that children under the age of 12 did perform more random acts of kindness than the older children. So she feels really confident that this effect is real.

A researcher in England investigated people's responses to sarcastic remarks. He had two groups of participants that came into his research laboratory to participate in the study. Another researcher acted as a confederate to say sarcastic remarks to the participants, but the participants believed that this person was another participant. Participants were divided into two groups, both groups were asked to respond as they naturally would to the sarcastic remark. One group was given the additional task of responding with a sarcastic remark and the other groups was told to respond with a remark that was not sarcastic. The researcher took physiological measurements of the participants (e.g., skin conductance, heart rate, etc.) along with recording the participants' responses (e.g., content, voice pitch, speaking rate, etc.). All of these measures were used to investigate if it is more natural to respond to sarcasm with sarcasm, or with explicit language. The researcher found that people who had to respond with a non-sarcastic remark showed more signs of anxiety and other forms of distress. So if someone is sarcastic with you, be sarcastic right back, it's healthier!

Appendix P – Far Transfer Task Posttest (Studies 3 & 4)

Please read over the experiment carefully and take your time responding to each choice. Click on those elements that you feel are a problem and leave blank those elements that you think are not a problem. In each experiment there may be no problem, one problem, or more than one problem.

___ Construct Validity
___ Control Group
___ Correlational Study
___ Experimenter Bias
___ Generalizability
___ Measurement Sensitivity
___ Random Assignment
___ Replication
___ No Problem

Have you ever used an automatic cleaner for your bathroom? Well now there's one for your kitchen floor too! An independent research team tested the effectiveness of this product. 10 researchers each found 6 dirty kitchens. They used the new product on 3 kitchens and used traditional cleaning products (i.e., a mop) on the other 3. The kitchens were cleaned daily for one week, then researchers evaluated each one. The 30 kitchens cleaned with the automatic cleaning product were rated as 50% cleaner by the researchers via a visual examination of each floor. So the independent researchers concluded that the automatic cleaning product was a really great cleaner. The company that made the cleaner decided to try a second study where they compared their product to the top competitor. The company recruited 60 people for the second study. The company decided to let people choose whether to use their new product or their competitor's cleaning product so they could also get information about the "shelf appeal" of their product. Luckily, it was a pretty even split (32 and 28, respectively). The new participants were given instructions to do the same type of cleaning that the independent research team did before. At the end of one week, the same researchers came and evaluated the cleanliness of each kitchen in the same way as the first study. This second study found that the automatic clear was superior to the top competitor's product. So now the company feels very confident that this product will be a huge success.

A trainer developed a new program that will have your dog well behaved over night. In past studies people have studied two training types: traditional and hypnosis. But no one has tried combining the two training types until now. Here's how the combined method works. Dogs go through a session on basic commands (traditional training). Then while the dogs sleep a tape is played that says the command, says the correct behavior, and gives praise (hypnosis). The trainers tested this out by having three comparison groups: new training (traditional + hypnosis), traditional, and do nothing. The dogs completed each training type under the supervision of the lead trainer, who is often referred to as the "Dog Whisperer" because he seems to have almost magical skills at getting dogs to behave. At the completion of each training type the dogs were evaluated on their behavior by trainers from a different kennel. The dogs that did the new program (traditional + hypnosis) did the best and it took less than 24 hours! The trainer claims that the hypnosis is what really makes the difference. So he is putting out a book and a video that will teach you how to get your dog trained over night. Trainers at a few other kennels heard about this program and decided to try it at their kennel. They all bought the book and video and followed all of the instructions very carefully. The new program wored at some kennels and didn't work as well at other kennels. The original trainer, the "Dog Whisperer," attributed these findings to some of the trainers not following the instructions carefully enough. So it might not be perfect, but if you need your dog trained this seems like the way to go!

Have you ever wanted to be funnier? Well the producers of a stand-up show have started a comedy class. Before advertising on T.V., they ran an study with 50 volunteers from a near by office building to participate in their study. Half of the people took their class for 1 week and half took a pottery class for 1 week. All of the participants attended class for 1 hour 3 times a week. All of the participants did 5-minutes of stand-up before and after taking either the comedy or pottery class. The tapes were evaluated in two ways. The first measure used was the number of jokes that the person told while on stage. Second, a trained therapist watched the video and rated the person on the Colenwald Public Speaking Anxiety Scale. The idea was that after the training the person should feel more confident while on stage, and therefore show less anxiety. The researchers found that all of the participants significantly improved in their stand-up abilities. To make sure that the class really worked the researchers ran another study. In the second study they used people who had auditioned for a spot in their stand-up show. These participants did everything the same as in the first study; except instead of taking the class 3 times in one week, they took the class once a week for 3 weeks. But this time they did not find that the class worked. So the researchers concluded that the first results were just a fluke and the class doesn't help people to be funnier.

Appendix Q – Design-A-Study Task Posttest (Studies 3 & 4)

Claim: Teachers always tell their students that it is better to study a little bit of the course material each day, rather than try to cram all of the studying into the night before the exam.

Imagine that you are going to design a study to test this claim. The questions below will address some of the important decisions that you will have to make when designing and running your study (e.g., independent variables, dependent variables, etc.). For each question, select the best answer choice. It is alright if you don't know the correct answer, try your best to answer each question.

1. There are lost of ways to measure the impact of students' study habits on learning. Which measure below has the greatest construct validity?
    a. Number of correct responses on a test
    b. Number of correct responses on a test and quality of study guide completion
    c. Rate of correct responses (number of correct answers divided by amount of time to answer each question) on a test
    d. Time taken to complete a test
2. In this study you are testing which type of study behavior is best for students. When you run your study, which of the following groups would be best to compare?
    a. Cramming the night before, studying small amounts of the material throughout the semester, and class attendance
    b. Cramming the night before and studying small amounts of the material throughout the semester
    c. Cramming he night before, studying small amounts of the material throughout the semester, and studying nothing at all
    d. Cramming the night before, cramming two hours before, and studying small amounts of the material throughout the semester
3. How would participants be put into the different conditions?
    a. Ask participants which type of studying (cramming the night before or studying small amounts throughout the semester) they do so the groups are as realistic as possible
    b. Use the order of participants walking in to the classroom on the first day of classes (e.g., first half will cram the night before and second half will study small amounts throughout the semester)
    c. Divide participants based on prior test scores into high performing and low performing groups
    d. When participants walk into the classroom, flip a coin and if it is heads the participant crams the night before and tails the participant will study small amounts throughout the semester

4. Let's say that you disagree with teachers and believe that cramming the night before a test is just as good as studying small amounts throughout the semester. When designing your study, which would be the best way to make sure your opinion did not bias the results?
    a. Have a second person that is not involved with your study deliver instructions to half of the participants and have a third person that is not involved with your study deliver instructions to the other half
    b. Have a second person that is not involved with your study deliver instructions to all participants
    c. Do not give participants any instructions, simply tell them to cram the night before or study throughout the semester
    d. Have a second person that believes the teachers' claim is correct run half of the participants and run half of the participants yourself
5. A past study tested this same claim about study habits with college students learning about Newtonian physics. Which study listed below would be the best replication study for your to conduct?
    a. Use the same college students from the first study
    b. Use the same college students from the first study, but make sure to assign them to a different condition in the replication study
    c. Use college students learning about English literature
    d. Use college students at the same university learning about Newtonian physics
6. Which of the following research settings will allow you the most confidence that your results will generalize and have confidence in the accuracy of your findings?
    a. Controlled research laboratory that removes other influential variables during studying and test taking
    b. Controlled research laboratory only for the testing portion of the study
    c. Allow participants to study in an environment of their choosing and to take the test in a similar environment (like a take home test)
    d. Allow participants to chose if they want to come into the laboratory to study and take the test or choose their own environment

Appendix R – Zohar & Nemet (2002) Coding Scheme

TABLE 2
The Sample Argument Coded Using Zohar and Nemet's Analytic Framework

Component of the Argument	Code	Scientific Knowledge
I think. . . all objects in the same surroundings become the same temperature even if an object produces its own heat energy.	Claim	Not coded
This is true because on the lab that we did all the temperatures were in their 20s which proves that the room temperature changes the objects to the same as the room.	Relevant justification	Correct scientific knowledge
Therefore, even though they may feel different, the objects are actually within a few degrees of each other.	Relevant justification	Incorrect scientific knowledge

*Table reprinted from Sampson & Clark (2008)

Appendix S – IRB Approval

THE UNIVERSITY OF MEMPHIS

Institutional Review Board

To: Sidney K. D'Mello
Computer Science

From: Chair, Institutional Review Board
for the Protection of Human Subjects
Administration 315

Subject: Promoting, Detecting, and Scaffolding Confusion and Cognitive Disequilibrium During Complex Learning (E09-278)

Approval Date: June 8, 2009

This is to notify you that the Institutional Review Board has designated the above referenced protocol as exempt from the full federal regulations. This project was reviewed in accordance with all applicable statutes and regulations as well as ethical principles.

When the project is finished or terminated, please complete the attached Notice of Completion and send to the Board in Administration 315.

Approval for this protocol does not expire. However, any change to the protocol must be reviewed and approved by the board prior to implementing the change.

Chair, Institutional Review Board
The University of Memphis

Dr. A. Graesser

Appendix T – Induction × Confusion × Intervention × Regulation Effort Interaction Results from Study 3

Claim quality models were significant when participants were either not confused and had low regulation effort or were confused and had high regulation effort. This dichotomy of cases follows an expected pattern. If participants are not confused than there is no need to put forth more effort during the confusion regulation task because there is no confusion to resolve, whereas the opposite is true in the confused and high regulation effort cases.

The findings from the not confused and low regulation effort cases suggest that claim quality was most impacted by the combination of the Convince Only condition with each induction condition. Specifically, participants in the Convince Only condition were more likely to present a correct claim when in the False-True condition (B = 2.36) but less likely in the True-False condition (B = 60.2) compared to the True-True condition ($\chi^2(2) = 10.3$, $p = .006$). In addition, when participants were in the True-True condition those in the Convince Only condition were less likely to present a correct claim than both the Convince then Read (B = 10.6) and Convince while Read conditions (B = 4.42), with the Convince the Read condition also more likely to present a correct claim than the Convince while Read condition (B = 6.22, $\chi^2(2) = 7.59$, $p = .022$). In contrast, when in the False-True condition participants in the Convince Only condition were more likely to present a correct claim than those in both the Convince then Read (B = 4.63) and Convince while Read conditions (B = 4.38, $\chi^2(2) = 6.62$, $p = .037$).

It is interesting that the combination of False-True induction condition and Convince Only intervention condition were likely to produce a correct claim when learners were not confused and had low regulation effort. In previous experiments participants have been found to generally agree with the tutor agent, who is incorrect in the False-True condition (D'Mello et al., 2014; Lehman et al., 2013), and this would most likely lead to an incorrect claim unless participants changed their opinion during the argument construction process. In addition, the Convince Only condition does not provide a resource to correct errors and misconceptions, which makes it seem unlikely that participants would change to the correct opinion. This finding is then somewhat anomalous.

The findings from the confused and high regulation effort cases were also somewhat perplexing. Participants in the Convince while Read condition were more likely to make a correct claim when in the no-contradiction control condition compared to both experimental conditions ($\chi^2(2) = 5.08$, $p = .079$, True-False: B = 1.62, False-True: B = 1.44). In addition, when participants were in the True-False condition, those in the Convince Only (B = 1.61) and Convince then Read conditions (B = 1.30) were more likely to present a correct claim than Convince while Read condition ($\chi^2(2) = 5.08$, $p = .079$). These findings suggest that participants were less likely to present a correct claim when they were successfully confused by the presentation of a contradiction, were asked to construct an argument, were provided with a resource to successfully resolve the contradicting opinions, and put forth more effort to resolve their confusion. In other words, this pattern of findings was the opposite of the predictions based on cognitive disequilibrium (Festinger, 1957; Graesser et al., 2005; Piaget, 1952) and impasse-driven theories of learning (Brown & VanLehn, 1980; VanLehn et al., 2003).

A presence score of 1 represents an argument that contains either a claim or evidence, but does not address the quality of the claim or evidence. The experimental conditions performed differently when participants were not successfully confused, had low regulation effort, and were in the Convince while Read condition ($\chi^2(2) = 4.51$, $p = .105$). When in the False-True condition participants were more likely to have a presence score of 1 than the True-True condition (B = 1.68), whereas the True-False condition was less likely to have a score of 1 (B = 1.24). When participants were in the True-False condition they were also less likely to have a presence score of 1 when they were confused, had low regulation effort, and were in the Convince Only condition ($\chi^2(2) = 5.74$, $p = .057$, B = 1.31). This pattern suggests that the True-False condition constructed arguments that were of an all or none nature. The arguments either contained both a claim and evidence or neither. However, when participants were confused, had high regulation effort, and were in the True-True condition those in the Convince Only (B = 1.65) and Convince then Read conditions (B = 1.00) were more likely to have a presence score of 1 than the Convince while Read condition ($\chi^2(2) = 4.76$, $p = .093$). It does not appear then that there is a simple pattern of events to explain a presence score of 1.

The semantic match score findings generally revealed that the False-True condition had a lower semantic match score than the True-True condition, regardless of the situation (i.e., case), whereas the intervention conditions greatly varied based on the situation. Participants had a lower match score to the ideal response when in the False-True condition compared to the True-True condition when they were not confused, had low regulation effort, and were in the Convince Only condition ($F(2,936) = 5.59$, $p = .007$, B = .295); were not confused, had high regulation effort, and were in the Convince while Read condition ($F(2,936) = 3.95$, $p = .028$, B = .319); and were confused, had low regulation effort, and were in the Convince while Read condition ($F(2,936) = 3.40$, $p = .041$, B = .200). However, participants did have a higher match score when in the False-True condition when they were confused, had high regulation effort, and were in the Convince Only condition ($F(2) = 3.02$, $p = .055$, B = .113). Participants also had higher match scores when in the True-False condition in this case (B = .170).

As mentioned previously, the intervention conditions performed differently in terms of semantic match score based on the situation. However, it was the case that the Convince Only condition generally had a higher match score than both the Convince then Read and Convince while Read conditions. The Convince Only condition had higher match scores than the Convince then Read condition when participants were confused, had low regulation effort, and were in either the True-True ($F(2,936) = 2.59$, $p = .084$, B = .232) or True-False conditions ($F(2,936) = 2.49$, $p = .095$, B = .185) and had higher match scores than the Convince while Read condition when participants were confused, had low regulation effort, and were in the False-True condition ($F(2,936) = 2.34$, $p = .103$, B = .176). The Convince then Read condition had higher match scores than the Convince while Read condition when participants were confused, had low regulation effort, and were in the False-True condition as well (B = .185). In addition, the Convince while Read condition had higher match scores than the Convince then Read condition when participants were confused, had low regulation effort, and were in the True-False condition (B = .149). The Convince while Read condition only had higher match scores than both the Convince Only (B = .290) and Convince then Read conditions (B = .167)

when participants were not confused, had high regulation effort, and were in the True-True condition ($F(2,936) = 4.61$, $p = .016$).

These findings were surprising for two reasons. First, the task of constructing an argument was identical in the Convince Only and Convince then Read conditions; therefore, a difference between the two conditions was not expected. Second, participants in the Convince Only condition were not provided with the explanatory text as a resource during argument construction as those in the Convince while Read condition were. The availability of the explanatory text was expected to aid in constructing an overall higher quality argument. However, the current findings did not support this prediction.

Appendix U – Induction Condition Differences for the Induction × Confusion × Intervention × Regulation Effort Interaction for the Far Transfer Task in Study 3

The induction condition differences for the induction × confusion × intervention × regulation effort interaction for the far transfer task revealed the circumstances under which the experimental conditions outperformed the no-contradiction control. When participants were confusion, had high regulation effort, and were in the Convince Only condition they performed better when in both experimental conditions than the no-contradictions control condition ($\chi^2(2) = 7.93$, $p = .019$; True-False: B = 1.14, False-True: B = 1.93). In addition, participants performed better when in the True-False condition when they were confused, had low regulation effort, and were in the Read Only condition ($\chi^2(2) = 4.54$, $p = .104$; B = 1.34), while they performed better when in the False-True condition when they were not confused, had high regulation effort, and were in the Convince then Read condition ($\chi^2(2) = 11.3$, $p = .004$; B = 8.13). These findings provide information as to the most beneficial combinations of induction and intervention conditions when participants were not in the overall most effective intervention condition (Convince while Read). It appears that when only one regulation task is presented the successful induction of confusion is needed for the contradicting information conditions to be effective.

Appendix V – Induction Condition Differences for the Induction × Confusion × Intervention × Regulation Interaction for the Near Transfer and Design-A-Study Tasks in Study 4

There were three main findings when induction condition differences were investigated for the near transfer task. The first finding was that participants did worse when in the False-True condition compared to the True-True condition (B = 3.85) when they were not confused, had high regulation effort, and were in the Confusion Specific condition ($\chi^2(2) = 8.74$, $p = .013$). This finding is somewhat puzzling because although the participants were not confused, they were somehow motivated to put in more effort during the regulation task. Although there is ostensibly not an impasse to resolve, it seems unlikely that the False-True condition would perform worse than the True-True condition. In contrast, the second main finding revealed that participants did better when in the True-False condition compared to the True-True condition when they had low regulation effort and were in the Confusion Specific + Motivation condition, regardless of whether confusion was successfully induced ($\chi^2(2) = 4.67$, $p = .097$, B = 1.18) or not induced ($\chi^2(2) = 15.5$, $p < .001$, B = 12.2). This finding is also somewhat perplexing due to the fact that the Confusion Specific + Motivation intervention did not successfully motivate participants (low regulation effort), the participants were still able to perform well on the near transfer task. The third main finding was the circumstances under which participants performed worse when in the True-False condition compared to the True-True condition. Participants did worse when in the True-False condition when they were not confused, had low regulation effort, and were in the General Motivational Statement ($\chi^2(2) = 7.12$, $p = .028$, B = 2.90) and were confused, had high regulation effort, and were in the Tutor Attribute + Motivation condition ($\chi^2(2) = 7.13$, $p = .028$, B = 1.41). This pattern suggests that a general motivational statement that does not address participants' attributions does not motivate participants to put in more effort and learn the concept more deeply.

The significant models for induction condition differences for the design-a-study task generally revealed the conditions under which the experimental conditions performed less well than the no-contradiction control condition, with one exception. As in the near transfer task analyses, when participants were not confused, had low regulation effort, and were in the General Motivational Statement condition, participants performed less well when in the True-False condition than the True-True ($\chi^2(2) = 6.30$, $p = .043$; B = 1.99) and the same pattern was found for the False-True condition in the present analyses (B = 1.31). Participants also performed less well when in both experimental conditions when they were not confused, had high regulation effort, and were in the Confusion Specific + Motivation condition ($\chi^2(2) = 6.95$, $p = .031$; True-False: B = 1.83, False-True: B = 3.31) and when they were confused, had high regulation effort, and were in the Material Attribute + Motivation condition ($\chi^2(2) = 11.7$, $p = .003$; True-False: B = 2.37, False-True: B = 1.56). This pattern generally suggests that the contradictory information conditions were not beneficial to performance on the design-a-study task. There was one exception to this pattern. When participants were in the True-False condition they performed better than the True-True condition when they were confused, had high regulation effort, and were in the Tutor Attribute + Motivation condition ($\chi^2(2) = 3.90$, $p = .071$, B = 1.30). This finding was the opposite of the pattern

that emerged in the near transfer study. In this instance the finding is more intuitive. Participants were successfully confused by the presentation of contradictory information and were motivated to put in more effort during the confusion regulation task. It is not clear why shifting the causal attribution of participants' confusion to the tutor agent's explanation was particularly effective in this circumstance. In order to fully explain why attributional shifts of different varieties and attributional retraining do and do not work in certain circumstances may necessitate a more complete understanding of the participants' attributions prior to the intervention.