

# Operation ARIES!: Methods, Mystery, and Mixed Models: Discourse Features Predict Affect in a Serious Game

CAROL M. FORSYTH  
The Institute for Intelligent Systems  
The University of Memphis  
Memphis, TN 38152  
[cmfrsyth@memphis.edu](mailto:cmfrsyth@memphis.edu)

and

ARTHUR C. GRAESSER, PHILIP PAVLIK JR. AND ZHIQIANG CAI  
The Institute for Intelligent Systems  
The University of Memphis  
Memphis, TN 38152  
[graesser@memphis.edu](mailto:graesser@memphis.edu), [ppavlik@memphis.edu](mailto:ppavlik@memphis.edu), [zcai@memphis.edu](mailto:zcai@memphis.edu)

and

HEATHER BUTLER AND DIANE F. HALPERN  
Claremont McKenna College  
Claremont, CA, 91711  
[hbutler@cmc.edu](mailto:hbutler@cmc.edu), [diane.halpern@claremontmckenna.edu](mailto:diane.halpern@claremontmckenna.edu)

and

KEITH MILLIS  
Northern Illinois University  
DeKalb, IL 60115  
[kmillis@niu.edu](mailto:kmillis@niu.edu)

---

Operation ARIES! is an Intelligent Tutoring System that is designed to teach scientific methodology in a game-like atmosphere. A fundamental goal of this serious game is to engage students during learning through natural language tutorial conversations. A tight integration of cognition, discourse, motivation, and affect is desired to meet this goal. Forty-six undergraduate students from two separate colleges in Southern California interacted with Operation ARIES! while intermittently answering survey questions that tap specific affective and metacognitive states related to the game-like and instructional qualities of Operation ARIES!. After performing a series of data mining explorations, we discovered two trends in the log files of cognitive-discourse events that predicted self-reported affective states. Students reporting positive affect tended to be more verbose during tutorial dialogues with the artificial agents. Conversely, students who reported negative emotions tended to produce lower quality conversational contributions with the agents. These findings support a valence-intensity theory of emotions and also the claim that cognitive-discourse features can predict emotional states over and above other game features embodied in ARIES.

Key Words: motivation, Intelligent Tutoring Systems, serious games, discourse, emotions, scientific inquiry skills, ARIES

---

## 1. INTRODUCTION

The goal of this study is to identify cognitive-discourse events that predict the affective valences (positive, negative) that learners experience in a serious game called Operation ARIES! [Millis et al. in press]. The game teaches students how to reason critically about scientific methods. The affective valences are reflected in self-reported metacognitive and meta-emotional judgments that college students provided after each of the

21 different lessons. Our expectation was that the cognitive-discourse events that reflect the students' academic performance on critical scientific thinking (i.e. the serious subject matter at hand) should be an important predictor of their emotional experience regardless of the impact of the narrative and other features in the game environment. However, our assumption may be incorrect. It may be instead that the other game aspects reign supreme and mask any impact of targeted cognitive-discourse factors on the emotions that occur during the learning experience.

The narrative features of games are known to have an influence on the students' emotional experience [Gee 2003; McQuiggan et al. 2010; Vorderer and Bryant 2006]. For example, there is the dramatic theme in which a heroic protagonist saves the world from destructive forces, a mainstay of movies, comic books and video games for decades. That theme exists in *SpyKids*, a movie that draws many children to theaters. In science-fiction novels, teenagers fantasize about being the hero who saves the world and wins the love of the pretty girl or handsome guy. Villains in these stories range from a single nemesis, such as in the comic book *Spiderman*, to groups of extra-terrestrials, such as "The Covenant" and "The Flood" in the science-fiction video game franchise "Halo" [Bungie.net]. These artificial worlds can be so engaging that gamers often use an acronym with negative connotations to describe real life (i.e. "IRL" stand for "in real life"). Movies and video games make billions of dollars for developers by allowing a single person to become immersed in a fantasy world where he or she is able to overcome seemingly in-conquerable antagonists and obstacles. It is not surprising that such narratives have an effect on players. Research in cognitive science and discourse processing has established that narratives have special characteristics [Bruner 1986; Graesser et al.1994] that facilitate comprehension and retention compared with other representations and discourse genres [Graesser and Ottati 1996].

Narrative is a feature that propels many serious games [Ratan and Ritterfeld 2009] and therefore a subject of interest for researchers. For example, *Crystal Island* [McQuiggan et al. 2010; Spires et al. 2010] is a serious game teaching biology that has been used to investigate narrative, emotions, and learning. The tracking of emotions within this system shows that students tend to become more engaged when the narrative is present compared with the serious curriculum alone. This engagement leads to increased learning [McQuiggan et al. 2008; Rowe et al. 2011]. However, another line of research [Mayer in press] suggests that the narrative component within the game may be distracting and contribute to lower learning gains than the pedagogical features alone. Although fantasy is fun and exciting, it runs the risk of being a time-consuming activity that can interfere with educational endeavors [Graesser et al. 2009; Mayer and Alexander 2011; O'Neil and Perez 2008]. The verdict is not in on the relationship between narrative and learning, but the impact of narrative on emotions clearly affects learning experiences.

Aside from narrative, there are other features within a learning environment that are likely to have an effect on emotions experienced by the student. It is conceivable that cognitive, discourse, and pedagogical components of serious learning can influence the affective experience [Baker et al. 2010; D'Mello and Graesser 2010; in press; Graesser and D'Mello 2012]. Research is needed to disentangle these aspects of a serious game from the narrative aspects in predicting learner emotions. The cognitive, discourse, and pedagogical aspects may potentially have a positive, negative, or non-significant impact on emotions. For example, learning academic content is considered quite the opposite of fun for many students. Instead of focusing on their uninteresting school activities, students are

prone to spend many hours playing video games and immersing themselves in worlds of fantasy [Ritterfeld et al. 2009]. The cognitive achievements in the mastery of serious academic content would have little or no impact on the affective experience in a serious game if the narrative game world robustly dominates. In essence, games and serious learning launch two separate psychological mechanisms, with very little cross talk, and the narrative game aspect dominates. An alternative possibility is that the cognitive, discourse, and pedagogical mechanisms still play an important role in determining emotional experiences over and above the narrative game aspects.

## 2. OVERVIEW OF OPERATION ARIES: A SERIOUS GAME ON SCIENTIFIC CRITICAL THINKING

Millis and colleagues [2011] developed the serious game *OperationARIES!* with the hope of engaging students and teaching them the fundamentals of scientific inquiry skills. The game will be referred to as ARIES, an acronym for *Acquiring Research Investigative and Evaluative Skills*. The storyline embedded in ARIES contains narrative elements of suspense, romance, and surprise. The game teaches scientific inquiry skills while employing the traditional theme of the single player saving the world from alien invaders threatening eminent doom.

Scientific reasoning was selected for ARIES because of the contemporary need for citizens to understand scientific methods and critically evaluate research studies. This is an important skill for literate adults in both developed and developing countries. When citizens turn on the television or log on to the internet, they are inundated with research reports that are quite engaging, but based on poor scientific methodology. One notable example in both popular and academic circles is the inference of causal claims from correlational evidence [Robinson et al. 2007]. For example, a recent study published the finding that drinking wine reduces heart risk in women. The study was correlational in nature but a causal claim was made. Perhaps it is true that drinking red wine is *associated* with reduced heart risk in women. But how can scientists legitimately claim that drinking red wine *causes* reduced heart risk without the use of random assignment to groups? Perhaps women who drink red wine also exercise more or are less likely to smoke, two more direct causes for reduced heart attacks, whereas there is no causal link between wine consumption and longevity. Most people are unable to distinguish true from false claims without training on scientific reasoning.

ARIES teaches students how to critically evaluate such claims by participating in natural language conversations with pedagogical agents in an adaptive Intelligent Tutoring System. ARIES has three modules: a Training module, a Case Study module, and an Interrogation module. It is the Training module that is under focus in the present study so we will not present details about the Case Study and Interrogation modules. In the Training Module, students learn 21 basic concepts of research methodology, including topics such as control groups, causal claims, replication, dependent variables, and experimenter bias. The Case Study and Interrogation modules apply these 21 concepts to dozens of specific research case studies in the news and other media.

The Training module has 21 chapters that are associated with each of the 21 core concepts. For each chapter, the students completed four phases:

Phase 1. Read a chapter in a book that targets the designated core concept (or topic).

Phase 2. Answer a set of multiple choice questions about the topic.

Phase 3. Hold a conversation about the topic with two conversational agents in a *dialog*, as discussed below. The students' conversational contributions serve as our predictor variables.

Phase 4. Give ratings on metacognitive and meta-emotional questions that are designed to assess the students' affective and metacognitive impressions of the learning experience. Such measures served as our criterion variables.

ARIES has an internal architecture in the Training module that is similar to that of AutoTutor, an Intelligent Tutoring System (ITS) with mixed-initiative natural language conversations. AutoTutor teaches students computer literacy and physics using Expectation-Misconception centered tutorial dialog [Graesser et al.2008; Graesser et al. 2004; VanLehn et al. 2007]. The pedagogical methods used in AutoTutor have shown significant learning gains comparable to expert human tutoring [Graesser et al. 2004; VanLehn et al. 2007]. This type of tutoring conversation revolves around a conversational pedagogical agent that scaffolds students to articulate specific expectations as part of a larger ideal answer to a posed question. In order to accomplish this goal, the artificial teacher agent gives the students *pumps* (i.e., "Um. Can you add anything to that?"), appropriate *feedback* ("Yes. You are correct", or "not quite"), *hints*, *prompts* to elicit particular words, misconception *corrections*, and *summaries* of the correct answer. The system interprets the students' language by combining Latent Semantic Analysis [LSA, Landauer et al. 2007], regular expressions [Jurafsky and Martin 2008], and weighted keyword matching. LSA provides a statistical pattern matching algorithm that computes the extent to which a student's verbal input matches an anticipated expectation or misconception of one or two sentences. Regular expressions are used to match the student's input to a few words, phrases, or combinations of expressions, including the key words associated with the expectation. For example, the expectation to a question in the physics version of AutoTutor is, "The force of impact will cause the car to experience a large forward acceleration." The main words here are "force," "impact," "car", "large," "forward," and "acceleration." The regular expressions allow for multiple variations of these words to be considered as matches to the expectation. Curriculum scripts specify the speech generated by the conversational agent as well as the various expectations and misconceptions associated with answers to difficult questions. A more in depth summary of the language processing used in AutoTutor is explained later in this paper as it applies to ARIES.

ARIES takes the pedagogy and principles of AutoTutor to a new level by incorporating multiple agent conversations in a game-like environment. ARIES holds three-way conversations (called dialogues) with the human student, who engages in interactions with a tutor agent and a peer student agent. These conversations with and between the two agents are not only used to tutor the student but also to discuss events related to the narrative presented in ARIES. As with AutoTutor, the ARIES system has dozens of measures that reflect the discourse events and cognitive achievements in the conversations. The log files record the agents' pedagogical discourse moves, such as feedback, questions, pumps, hints, prompts, assertions, answers to student questions, corrections, summaries, and so on. These are the discourse events generated by the agents, as opposed to the human student. The log file also records cognitive-discourse variables of the human student, such as their verbosity (number of words), answer quality (semantic match scores to expectations versus misconceptions), reading times for texts, and so on. These measures are a mixture of the cognitive achievements and discourse events so we refer to them as cognitive-

discourse measures. The question is whether these cognitive-discourse measures that accrue during training predict the students' self-reported affect (i.e., emotional and motivational metacognitive states) after lessons are completed with ARIES. Data mining procedures were applied to assess such predictive correlational relationships. It should be emphasized that the goal of the present study is on predicting these self-reported affective impressions, not on learning gains.

The cognitive-discourse measures are orthogonal to the other game-like features of ARIES, such as the narrative storyline and multimedia presentations. An alternative possibility is that the flashy dimensions of the game and narrative would account for the affective experience, thus occluding any cognitive-discourse features measured during the pedagogical conversations. ARIES was designed to engage the student through a complex storyline presented through video-clips, e-mails, accumulated points, competition, and other game elements in addition to the conversational dialogues. Within the Training module alone, the storyline allows for role-play, challenges, fantasy and multi-media representations which are all considered characteristics that amplify the learning experience in serious games [Huang and Johnson 2008]. The embedded storyline includes interactive components with the student which is a mainstay of games [Whitton 2010]. In addition, the student is given some player control over the course of the Training module which may be an effective method of engaging students [Malone and Lepper 1987]. The developers took care to attempt to balance the game-like features with the pedagogical aspects to maximize the learning experience without over-taxing the cognitive system which can be a pitfall for serious game developers [Huang and Tettegah 2010]. However, the possibility must be considered that perhaps these game-like features end up being so strong that cognitive-discourse features pale in comparison.

### 3. EMOTIONS DURING LEARNING

Contemporary theories of emotion have emphasized the fact that affect and cognition are tightly coupled rather than being detached or loosely connected modules [Isen 2008; Lazarus 2000; Mandler 1999; Ortony et al. 1988; Scherer et al. 2001]. However, the role of affect in learning of difficult material is relatively sparse and has only recently captured the attention of researchers in the learning sciences [Baker et al. 2010; Calvo and D'Mello 2010; Graesser and D'Mello 2012]. A distinction is routinely made between affective traits versus states, whether they be in clinical psychology [Spielberger and Reheiser 2003] or advanced learning environments [Graesser et al. 2012]. A *trait* is a persistent characteristic of a learner over time and contexts. A *state* is a more transient emotion experienced by a student that would fluctuate over the course of a game session. For example, some people have the tendency to be frustrated throughout their lives and adopt that as a trait. In contrast, the state of frustration may ebb and flow throughout a game or any other experience. The same is true for many other academic and non-academic emotions, such as happiness, anxiety, curiosity, confusion, boredom, and so on. Therefore, it is inappropriate to sharply demarcate emotions as being either traits or states because particular emotions can be analyzed from the standpoint of traits (constancies within a person over time and contexts) and states (fluctuations within an individual and context). The present study addresses affective states rather than traits during the course of learning.

One explanation for the effect emotions have during learning is the *broaden and build theory* [Frederickson 2001], which was inspired by Darwinian theory. The broaden and build theory is based on the idea that negative

emotions tend to require immediate reactions whereas positive emotions allow one to view the surroundings (broaden) and accrue more resources in order to increase the probability of survival (build). In a learning environment, positive emotions have been associated with global processing, creativity and cognitive flexibility [Clore and Huntsinger 2007; Frederickson and Branigan 2005; Isen 2008] whereas negative emotions have been linked to more focused and methodical approaches [Barth and Funke 2010; Schwarz and Skumik 2003]. For example, students in a positive affective state may search for creative solutions to a problem whereas one experiencing a negative affective state may only employ analytical solutions. This is one explanation of the effect of emotions on learning, but there are others.

Another explanation of the relationship between affect and learning considers the match between the task difficulty level and the student's skill base. The theoretically perfect task for any student would be one that is neither too easy nor too difficult but rather within the student's zone of proximal development [Brown et al. 1998; Vygotsky 1986]. When this optimal alignment occurs, students may experience the intense positive emotion of flow [Csikszentmihalyi 1990] in which time and fatigue disappear. There is some evidence that students are in this state when they quickly generate information and are receiving positive feedback [D'Mello and Graesser 2012; 2010]. For example, a student in the state of flow may make many contributions to tutorial conversations and remain engaged for an extended period of time. Conversely, a student who is in a negative emotional state may only make minimal contributions, experience numerous obstacles, and thereby experience frustration or even boredom.

Students' self-efficacy or beliefs about their ability to complete the task may also relate to emotional states experienced during learning. Specifically, students who appraise themselves to be highly capable of successfully completing a task are experiencing high self-efficacy. These students exert high amounts of effort and experience positive emotions [Dweck 2002]. Conversely, students who feel that failure is eminent will exert less effort, perhaps resulting in insufficient performance and negative emotions. However, beliefs of self-efficacy may dynamically change throughout a learning experience [Bandura 1997], thus acting as states rather than traits of the student. For example, a student may begin a task with a relatively high level of self-efficacy and then discover that the task is more difficult than expected, leading to a decreased level of self-efficacy. Students take into account previous and current performance when forming these transient beliefs about current states of self-efficacy.

Traits such as motivation and goal-orientation may add another dimension to the complex interplay between emotions and learning. Pekrun and colleagues [2009] suggest a multi-layered model of learning, motivation, and emotion. Pekrun's model, known as the control-value theory of achievement emotions, posits that the motivation or goal-orientation of the student is correlated with specific emotions during learning, such as enthusiasm and boredom. The student's goal-orientation can be focused on either *mastery* or *performance*. A student with mastery goal-orientation wants to learn the material for the sake of acquiring knowledge. This type of student sees value in learning simply for the sake of learning. In contrast, a student with performance goal-orientation wants to only learn enough information to perform well on an exam. Mastery-oriented students are more likely to achieve deep, meaningful learning, while the latter is likely to only acquire enough shallow knowledge to get a passing grade [Pekrun 2006; Zimmerman and Schunck 2008]. These two types of learners are likely to experience two vastly different emotional trajectories [D'Mello and Graesser 2012]. Mastery-oriented students might experience feelings

of engagement when faced with a particularly difficult problem, because difficult problems provide the opportunity to learn, grow, and conquer obstacles. They might even experience the aforementioned state of flow. That is, when an obstacle presents itself, the student experiences and resolves confusion and thereby maintains a state of engagement. On the other hand, performance-oriented students might feel anxiety, stress, or frustration when faced with a difficult problem, since performing poorly might threaten their grade in the class. When this type of student is faced with such problems, the student may experience confusion followed by frustration, boredom and eventual disengagement. The affective trajectories of the two separate goal-orientated traits of students results in an aptitude-treatment interaction, which is quite interesting but beyond the scope of this article to explore. The current study focuses on moment-to-moment affective states during the process of learning.

Several studies have recently investigated the moment-to-moment emotions humans experience during computer-based learning [Arroyo et al. 2009; Baker et al. 2010; Kapoor et al. 2007; Calvo and D'Mello 2010; Conati and Maclaren, 2009; D'Mello and Graesser 2012; Litman and Forbes-Riley 2006; McQuiggan et al. 2010]. Conati and Maclaren conducted studies on the affect and motivation during student interaction with an Intelligent Tutoring System called *PrimeTime*. These studies collected data in both the laboratory and school classrooms [Conati et al.2003; Conati and Maclaren 2009]. Emotions were detected by physical sensors which were recorded in log files along with cognitive performance measures. These studies resulted in a predictive model of learners' goals, affect, and personality characteristics, which are designated as the student models [Conati and Maclaren 2009].

Baker and colleagues [2010] investigated emotions and learning by tracking emotions of students while interacting in three very different computer learning environments: Graesser's AutoTutor (as referenced earlier), Aplusix II Algebra Learning Assistant [Nicaud et al. 2004], and a simulation environment in which students learn basic logic called *The Incredible Machine: Even More Contraptions* [Sierra Online Inc. 2001]. Baker et al. investigated 7 different affective states (i.e. boredom, flow, confusion, frustration, delight, surprise and neutral). Among other findings, the results revealed that students who were experiencing boredom were more likely to "game the system," a student strategy to trick the computer into allowing the student to complete the interaction without actually learning [Baker et al. 2006; Baker and deCalvalho 2008]. Frustration was expected to negatively correlate with learning, but it was low in frequency. Interestingly, confusion was linked to engagement in this study and has also been linked to deep level learning in previous studies [D'Mello et al. 2009]. Baker et al. [2010] advocated more research on frustration, confusion, and boredom in order to keep students from disengaging while interacting with the intelligent tutoring systems.

A series of studies examined emotions while students interacted with AutoTutor, which has similar conversational mechanisms as ARIES except that ARIES has 2 or more conversational agents and AutoTutor has only one agent. These studies conducted on AutoTutor investigated emotions on a moment-to-moment basis during learning [Craig et al. 2008; D'Mello et al. 2009; D'Mello and Graesser 2010; D'Mello and Graesser 2012 ; Graesser et al. 2008]. The purpose of these studies was to identify emotions that occur during learning and uncover methods to help students who experienced negative emotions such as frustration and boredom. As already mentioned, the most frequent learner-centered emotions in the computerized learning environments were confusion, frustration, boredom, flow/engagement, delight and surprise. These emotions have been classified on the dimension of valence,

or the degree to which an emotion is either positive or negative [Barrett 2007; Isenhower et al. 2010; Russell 2003]. That is, learner-centered emotions can be categorized as having either a negative valence (i.e., frustration and boredom), a positive valence (flow/engagement and delight), or somewhere in-between, depending on context (e.g., confusion and surprise).

Positively- and negatively-valenced emotions have been shown to correlate with learning during student interactions with the computer literacy version of AutoTutor [Craig et al. 2004, D’Mello and Graesser 2012]. Specifically, positive emotions such as flow/engagement correlate positively with learning, whereas the negative emotion of boredom shows a significant negative correlation with learning. Interestingly, the strongest predictor of learning is the affective-cognitive state of confusion, a state which may be either positive or negative, depending on the attribution of the learner. A mastery-oriented student may regard confusion as a positive experience that reflects a challenge to be conquered during the state of flow. A performance-oriented student may view confusion as a negative state to be avoided because it has accompanying negative feedback and obstacles. Available research suggests there may be complex patterns of emotions that impact learning gains rather than a simple relationship [D’Mello and Graesser in press; Graesser and D’Mello 2012]. Methods in educational data mining are expected to help us discover such relationships.

One motive for uncovering the complex array of emotions experienced during learning is to create an adaptive agent-based system that is responsive to students experiencing different emotions. Many systems with conversational agents have been developed during the last decade that have a host of communication channels, such as gestures, posture, speech, and facial expressions in addition to text [Atkinson 2002; Baylor and Kim 2005; Biswas et al. 2005; Graesser et al. 2008; Graesser et al. 2004; Gratch et al. 2001; McNamara et al. 2007; Millis et al. 2011; Moreno and Mayer 2004]. In order to provide appropriate feedback to the student, the system must be able to detect emotions on a moment-to-moment basis during learning. This requires an adequate model that detects if not predicts the students’ emotions.

The current study probed students’ self-reported impressions of their learning experiences with respect to emotions, motivation, and other metacognitive states after they interact with lessons in the Training module of ARIES. ARIES is an ITS similar to AutoTutor but it has game features and multiple agents. Unlike many of the AutoTutor studies on emotions, the current investigation has an added component of ecological validity because data were collected in two college classrooms. We did not want to disrupt the natural flow of classroom learning, so the students were not hooked up to technological devices that sense emotions. Additionally, the nature of the design included students working in dyads to complete interaction of ARIES. Students often work in pairs in classes so that a solitary student is not stuck for long periods of time.

The detection of emotions was based entirely on self-report measures collected after students completed each of the 21 chapters on research methodology. Self-report has been used in studies determining the validity of both linguistic [D’Mello and Graesser in press] and non-verbal affective features [Craig et al. 2004]. The survey questions did not ask students open-ended questions such as “how do you feel?” Instead, the questions asked students for levels of agreement with a set of statements that addressed specific aspects of the pedagogical and game-like features of ARIES, as will be discussed in the Methods section. The collection of ratings on multiple

statements that tap underlying constructs is the normal approach to assessing motivation and emotion in education [Pekrun et al. 2010]. According to attribution theories [e.g. Weiner 1992], people may be more likely to report an emotion when it is directed at an external object rather than oneself. This technique is used in order to decrease feelings of personal exposure and vulnerability, which hopefully yields more truthful answers. The interaction with ARIES provided many cognitive-discourse measures during learning which were explored as predictors of the students' self-reported affective states.

#### 4. DATA COLLECTION, EXPERIMENTAL DESIGN AND PROCEDURE

Data were collected from 46 undergraduates enrolled in Research Methods courses at two colleges in Southern California, an elite liberal arts college and a state school. The entire data collection occurred over a semester with students working on ARIES 2 hours per week. The experimental procedure consisted of first giving participants a pretest on knowledge of scientific methods, followed by interacting with ARIES across the three modules described earlier (Interactive Text, Case Study, Interrogation), and ending in a posttest on scientific methods. Each module took different amounts of time for completion: Training module (10 to 12 hours), Case Study (8-10 hours), and Interrogation Module (8-10 hours). During interaction with the Training module, students were grouped into pairs who alternated being passive and active players of the game. For example, participant A in the dyad typed input into the ITS in Chapter 1 while participant B observed; in Chapter 2, participant B actively interacted with ARIES while participant A observed. The active and passive students were allowed to chat about the material during the interaction with ARIES. The active versus passive learning was a within-subjects variable. All students completed a survey after each chapter in which they were asked to rate themselves on their affective and metacognitive impressions of the learning experience. The surveys probed their perceptions of the storyline, natural language tutorial conversations, how much they learned, and their emotions. Each student went on to complete the Case Study and Interrogation modules but there were no self-report questions during these modules. Finally, the students were given a post-test in order to gain a measure of total learning gains across all three modules. However, the present study focuses only on the self-report ratings as the criterion variables rather than the learning gains.

##### 4.1 Agent-Human Interaction

In the Training module, there was a series of 21 chapters, each of which was studied in 4 phases. These chapters of scientific methodology are about theories, hypotheses, falsifiability, operational definitions, independent variables, dependent variables, reliability accuracy and precision, validity, objective scoring, replication, experimental control, control groups, random assignment, subject bias, attrition and mortality, representative populations, sample size, experimenter bias, conflict of interest, making causal claims, and generalizability. For each chapter, students first read an eBook (Phase 1), then answered 6 multiple-choice (MC) questions (Phase 2), and then conversed with the two pedagogical agents in a triad (Phase 3) immediately after each of the final three MC items. Throughout this learning experience, the plot of aliens invading the world was continually presented through the speech of the artificial agents as well as e-mails sent to the human student. However, as discussed in the Introduction, these game-like and multimedia features were orthogonal to the cognition

and discourse reflected in the tutorial trialogs. A picture of the interface of the Training module can be found in Appendix A. Upon completion of each chapter, the students answered 14 questions about affective and metacognitive responses to the learning material as well as the storyline (Phase 4). This process occurred iteratively across the 21 chapters on scientific methodology.

Some additional details are relevant to Phases 1, 2, and 3. The student initially had a choice to either read the E-book (Phase 1) or immediately progress and take the MC questions (Phase 2). However, the vast majority of students elected to read the text. Each of the 21 chapters contained 3 sub-sections on the definition, importance, and an example of each specific research methodology topic. The MC questions each contained 2 questions about the definition, 2 questions about the importance, and 2 questions about the example. For example, the three types of questions evaluating the topic of operational definitions are “Which of the following statements best describes an operational definition?” as a definition question, “Which of the following statements best reflects why it is important to have operational definitions in a research study?” as an importance question, and “Which is NOT a good example of an operational definition?” as an example question. Students were classified into low, medium, and high mastery for each sub-section based on performance on the MC questions. If the student missed any of the 6 questions, he or she was required to re-read the E-book. Otherwise, the student progressed immediately into the next phase of tutorial dialog with the agents. Therefore, either after re-reading the E-book or simply taking the challenge test, the computer subsequently assigned the students to adaptive tutorial conversations (one for each sub-section, resulting in a total of 3 trialogs per chapter) with both the student and teacher pedagogical agents (Phase 3). The three modes consisted of *vicarious learning*, *standard tutoring*, and *teachable agent*. In the vicarious learning mode, students simply watched the teacher agent teach the student agent, but periodically the students were asked a YES/NO question to make sure they were attending to the conversation. At the other extreme, in the teachable agent mode, students taught the student agent. This was considered the most difficult trialog because the human served the role of a teacher. In the standard tutoring mode, the intermediate level, students participated in tutorial conversations with the teacher agent and the student agent whom periodically participated in the conversation. For the purposes of this paper, we focus only on the standard tutoring and the teachable agent mode because features of the students’ discourse can be analyzed. We do not consider this a serious limitation because the standard tutoring and teachable agent modes accounted for over 90% of the conversations that the students participated in due to performance on multiple-choice tests.

In the standard tutoring and teachable agent modes, the pedagogical conversation with the artificial agents was launched after the MC questions by asking an initial question to students. An example conversation can be found in Appendix B. The student-provided answer responses vary in quality with respect to matching a sentence-like expectation. For example, in Appendix A, one can see the expectation aligned with the initial question asking about why operational definitions are important. Specifically, the expectation is “Operational definitions are important because they allow a particular variable to be reliably recognized, measured and understood by all researchers.” If the student answered the initial question closely to the expectation, then the tutor gave positive feedback and moved forward in the conversation. However, most of the students’ answers were vague and did not semantically match the expectation at a high enough threshold. The tutor tried to get more out of the student with a

pump (“Tell me more.”), but that open-ended response may still not have been enough of a match. If students were unable to provide an accurate answer, then the agents scaffolded the students’ understanding by providing appropriate feedback, hints, prompts, and correcting misconceptions if necessary. For example, if students were unable to provide much information, they would first be given a hint. This is a broad question (e.g., What about X?) that should help the student recall or generate some relevant information about the topic. If students were still unable to provide a correct answer, the agents would give a prompt to elicit a specific word. For example, a prompt when trying to get a student to articulate the word “increase” might be, “The values on the variable do not stay the same but rather they what?” A prompt is easier to answer than a hint because it only requires a single word or phrase as an answer. When all else fails, ARIES simply tells the student the right answer with an assertion. The degree of conversational scaffolding that a student receives to get them to articulate an answer increases along the following ordering in discourse moves: pump → hint → prompt → assertion.

Once again, every question asked by the pedagogical agents has a corresponding expectation or ideal answer (the two terms are used interchangeably within ARIES). For example, a question requiring such an answer could be, “What aspects constitute an acceptable measure of the independent variable?” The expectation is “The dependent variables must be reliable, accurate and precise.” The human response to this answer is computed by a mixture of Latent Semantic Analysis [Landauer et.al 2007], regular expressions [Jurafsky and Martin 2008], and word overlap metrics that adjust for word frequency. The regular expressions correspond to three to five important words within the expectation. In the above example, the important words from this ideal answer are “dependent variable”, “must”, “reliable”, “accurate”, and “precise”. Throughout the conversation with the agents, the agents attempted to get the student to articulate these five words or close semantic associates determined by LSA. Semantic matches are to some extent computed by comparing the human input to the regular expressions that are implemented in ARIES. Regular expressions allow for more alternative articulations of an ideal answer. In the above example, the student needs to type “dependent variable” and “reliable” in the same utterance. The regular expression “(dv|\bdependent variable|measure|outcome).\*(reliab|consistent)” constrains the matching by forcing the student to type two words in the same utterance. It allows for alternate articulations including synonyms and morphemes of “dependent variable” and “reliable” to be accepted. The regular expression above also constrains the student from saying “independent variable” when the correct answer is “dependent variable.” In order to accept an expectation as being covered, the student must reach a .51 overlap threshold (including the regular expression match and the LSA match score – a cosine value) between the student’s language over multiple conversational turns and a sentence-like expectation. Latent Semantic Analysis has been successfully used as a semantic match technique in AutoTutor as well as in automated essay scoring [Graesser et al. in press; Landauer et al. 2007]. The method of computing semantic matches in ARIES is quite reliable. Specifically, semantic similarity judgments between two humans are not significantly different from the semantic matching algorithms implemented in the Training module of ARIES [Cai et al. 2011]. The semantic match measure served as another cognitive-discourse measure in the data mining analyses.

## 4.2 Survey Measures

Students were asked 14 questions after every chapter within ARIES. Therefore, each student filled out 21 surveys consisting of 14 questions each (294 answers total). Students responded to each question using a 7-point Likert scale (1= not at all to 7= a lot). Twelve of the measures were directed at the participant and two of the questions were directed at the participant's view of his/her partner. For example, a participant-directed question could be "How frustrating did you find the trialog to be?". Conversely, a partner-oriented question could be "How much do you think your partner is learning?". Self-report measures about one's self have high face validity and are routinely collected in educational research on motivation, affect, and metacognition [Pekrun et al. 2010]. However, peer judgments of affect and metacognition are not expected to be very valid because of the limited interactions between dyads. Therefore, for the purposes of categorizing individual affective and metacognitive states, this study only evaluated the self-directed questions. The survey questions included judgments of both the storyline and the trialogs. Questions about the storyline included likeability, quality, suspense, and curiosity. Trialog (or conversation) focused questions included measures of frustration, difficulty, and helpfulness.

## 4.3 Procedure

Participants began by taking a pretest that included multiple-choice and open-ended questions, each of which assessed prior-knowledge of research methodology. Next, they interacted with ARIES. Participants alternated between passive and active participants for every chapter. After completion of each chapter, participants were given the survey measures to rate and subsequently moved onto the next chapter. After completion of the Training module, participants completed the Case Study and Interrogation modules, followed by a posttest (including multiple-choice and open-ended questions). There were two versions of these tests (A and B); the pretest and posttest were counter-balanced (AB versus BA) and students were randomly assigned to these versions.

During the course of interacting with the Training module, the system logged over 150 specific measures, including string variables on students' verbal responses. All measures were recorded at the most fine-grained level possible. These included a score for each question on the multiple-choice tests, the number of words expressed by the student in each conversational turn, the type of trialog for each conversation, the match score between student verbal contributions and the expectation (for every level of the conversation including initial response, response after pump, after hint, and after the prompt). In addition to these measures, latency variables included the time spent on each trialog, the time each discourse move was introduced by either the student or the artificial agents, the time spent reading the book, and the start and stop time of each overall interaction and each chapter individually. String variables included the precise nature of each discourse move that was contributed by the artificial agents as well as the student, and the answer given by the student to the multiple-choice questions. By the end of the full experiment, we had upwards of 6000 measures obtained across the three modules. However, only the quantitative measures resulting from interaction with the Training module were used as predictors of the affective and metacognitive impressions. Qualitative measures would require additional computations of linguistic measures, such as syntactic complexity, given vs. newness, and word concreteness. Such additional computations would result in more inferential statistical tests performed, thus increasing the overall probability of Type I error.

## 5. ANALYSES AND RESULTS

Answers to the survey measures served as the criterion variables. Instead of having 12 independent ratings associated with the 12 questions, we performed a Principal Components Analyses (PCA) to reduce the 12 ratings to a smaller set of factors, called principal components. The cognitive-discourse measures that ended up being used were identified after mining dozens of indices that were collected during the student interaction with the Training module. After we converged on a small set of cognitive-discourse measures, we used mixed models and 10-fold cross validation to assess the impact of these measures over and above the distributions of scores among participants and among the 21 chapters.

We needed to take some steps in cleaning the data before conducting the analyses. Upon initial inspection of the data, we found one participant to have an extremely low negative value for a learning gains score. Therefore, we made the assumption that this student did not take the experiment seriously and deleted him/her from the dataset. Other participants were removed from the data set for various reasons, leaving a total N of 35. The reasons for deleting students included late registration and/or failure to follow teacher instructions, resulting in the lack of completion of survey measures or the Training module. Descriptive statistics on the variables also revealed a number of missing observations on the item (chapter) level. The participants took nearly a semester to complete interaction with ARIES because the full game lasts up to 22 hours. Whenever class was dismissed, students logged out of ARIES and upon returning, logged back into ARIES. Due to normal real-life circumstances, some participants missed some days and chapters. This resulted in missing observations, the majority of which were deleted from the dataset. Our dataset was reduced from 967 observations to 523 observations.

In our initial peak at the data, we examined first-order bivariate correlations between the logged measures and the survey ratings. We discovered that the ratings from the surveys (covering 12 questions) with the dozens of cognitive-discourse indices did not unveil readily detectable patterns, given the modest sample size of participants. Therefore, we needed to identify a smaller number of criterion components and a smaller number of cognitive-discourse predictors.

### 5.1 Principal Components Analysis (PCA) on Survey Ratings.

A principal component analysis (PCA) was conducted to reduce the 12 survey questions to a small number of meaningful components. A PCA is a type of exploratory factor analysis used to reduce a large number of variables into a smaller set of meaningful principal components after grouping individual scores into discrete components. An exploratory rather than confirmatory factor analysis was chosen because the researchers began the investigation undecided about the specific groupings of the affective and metacognitive reactions. The survey questions were originally constructed in order to obtain measures of student meta-emotional and metacognitive reactions to both the pedagogical and game-like aspects of ARIES. In the post hoc exploratory factor analysis, we discovered that the 12 survey measures (including both reactions to the pedagogical agents as well as the game-like aspects) grouped into cohesive principal components of positive valence, negative valence, and perceived learning. As an alternative, the

measures could have potentially been grouped into pedagogical vs. game-like features, but such a distinction was not substantiated compared to the three principal components.

The PCA was run on the matrix of results for the 35 participants, each with 21 topics, in violation of the normal requirement for independent observations. The researchers were justified in violating this assumption as the biasing of any factor scores were expected to be minimal because the 21 observations were temporally separate and individually contextualized within the ARIES software experiences. While this certainly introduces some risk in terms of PCA parameters such as overall variance accounted for, it still seemed implausible that the PCA method would fail to find the overall structure characterizing the principal components of student responding. Any bias due to change in the reporting by students across the 21 observations was expected to be small. For the purpose of the investigation, it seemed the best option to average all of the observations into the overall factors.

A PCA allows for iterations of rotations to be performed on the identified measures in order to find an optimal meaningful grouping of the data. In performing this analysis on the survey measures, the direct oblimin with Kaiser Normalization method was chosen over the varimax rotation because of anticipated inter-correlations in the Likert-survey questionnaires. The loading of the survey measures into the direct oblimin solution yielded three principal components that exceeded an Eigenvector value of 1 with minimal oblique components after rotation. Thus, the survey measures converged on three principal components. The total variance explained by the 3 principal components (PC's) was 75.5%. The variance ( $R^2$ ) accounted for by the three principal components was 52.5%, 13.9%, and 9.2%, respectively, when computing the total sum of squares. The list of factors can be seen in the order of variance explained in Table 1.

Table I: Principal Component Analysis of Survey Questions.

Survey Questions	PC1	PC2	PC3
Participant believes trialogs were frustrating		+(0.640)	
Chapter difficulty			+(0.933)
Quality of the chapter	+(0.558)	-(0.810)	
Amount participant felt learned		-(0.696)	+(0.570)
Quality of storyline	+(0.828)	-(0.581)	
Quality of ebook	+(0.629)	-(0.737)	
Participant liked storyline	+(0.837)	-(0.537)	
Participant thought the storyline was suspenseful	+(0.916)		
Participant felt surprised by storyline	+(0.906)		
Storyline evoked curiosity	+(0.917)		
Participant believes trialogs helped with learning		-(0.769)	
Quality of trialogs		-(0.827)	

*Note.* A plus (+) or minus (-) sign designates primary loading on a PC, in either a positive or negative direction.  
PC= Principal Component PC1= Positive affect PC2=Negative valence PC3= Perceived Learning

The three principal components could be readily interpreted as falling into the following three dimensions: *positive affective valence* (PC1), *negative affective valence* (PC2), and *perceived learning* (PC3). The first two principal components are well aligned with the valence-arousal model that is widely accepted in general theories of emotions

[Barrett 2006; Russell 2003]. The third principal component is a metacognition dimension that captures the students' judgment of learning as well as perceptions of chapter difficulty. One can infer that perceived difficulty of the chapter will likely have a negative correlation with how much the student feels he or she is learning. Students are notoriously inaccurate at judging their own learning [Dunlosky and Lipko 2007; Maki 1998], but they nevertheless do have such impressions that were captured by the third principal component (PC3). The subsequent analyses were conducted on the factor scores of these three principal components, each of which served as a criterion variable. The question is whether there are cognitive-discourse events in the trialogs that predict these principal components that capture affect and metacognition.

## 5.2 Correlations between Cognitive-Discourse Measures and Affect and Metacognitive Loadings

A correlation matrix was computed in order to get an initial grasp on which logged measures of cognition-discourse might best predict affect and metacognition loadings. Only quantitative measures, rather than qualitative measures, were used in this analysis in order to keep the number of statistical tests to a minimum and reduce the probability of Type 1 error. The correlations were performed on the item level rather than subject level because of the low number of participants as well as the variance within subjects. The researchers were confident in using this level of analysis because the Pearson correlations were calculated merely as a guide for creating predictive models on the item level, which would later be confirmed with cross-validation. In determining substantial correlations, a threshold of  $r > .2$  was used rather than statistical significance ( $p < .05$ ), which was lower than .2 as correlations were performed with many observations ( $N = 523$ ). The effect size proved to be more reliable when creating predictive models.

Using the above criterion, some of the conversational features appeared to be sufficiently correlated with the survey measures. We analyzed many measures of cognition and discourse; for each, there were scores that covered observations across 35 students, 21 chapters, and three trialogs per chapter. The three trialogs per chapter each addressed the definition, importance and example sub-section of the given topic. For each of the three sub-sections, the reader may recall that multiple levels of scaffolding are employed by the pedagogical agents including an initial question, pump, hint, and prompt. In the following analyses, the cognitive-discourse features including match score and words generated were only analyzed on the level of the student answer to the initial question, rather than input for each level of scaffolding. The unit of analysis allowed the maximum comparable number of observations for the predictor variables. Additional measures were obtained from the students' interactions with the 21 chapters in ARIES. The measures included scores for answers to each of the MC questions presented within a chapter, as well as performance scores for each sub-section of a chapter (including a single score ranging from 0-1 for the definition questions, the importance questions, and the example questions). Multiple time variables were analyzed specifically focusing on the time spent during tutorial conversations for each sub-section as well as reading the E-text chapter. Finally, the type of trialog (i.e. standard tutoring or teachable agent) and gender of the student were investigated as possible predictors of the three factors.

After inspecting 57 correlations with the principal components, we discovered three correlations that were relatively robust. The match score correlated with negative ratings ( $r = -.289$ ), the word count correlated with the positive ratings ( $r = .231$ ), and the reading time correlated with ratings of perceived learning ( $r = .250$ ). Match score

is a measure of the *quality* of the student's input whereas word count is a measure of the *quantity* of a student's generated input. The two conversational measures of match score and number of words were not significantly correlated with each other ( $r = -.069$ ). For all three principal components, variability across the 21 chapters was observed. On a finer-grained level, match score and words generated were investigated on the level of each sub-section for each chapter (definition, importance, example). However, the cumulative score across all three sub-sections was either equal to or greater than any sub-section for both word count and match score. In order to increase the generalizability of the results, the cumulative score was used to represent each variable. Though none of the additional measures reached a threshold of  $r > .2$ , they are reported in Table II for the sake of informing the reader.

In summary, correlational analyses revealed that the match scores of the trialog initial turn, word count and reading times appeared to have the most predictive correlations of the factor scores associated with the survey ratings. Match scores served as an index of the quality of the students' contributions and correlated with the negative valence component. The predictor of word count served as a measure of quantity of words generated by the student and was the highest correlate of the measure of positive valence. Finally, reading times were the highest correlate of perceived learning and were operationally defined as the duration of time between the student opening and closing the E-book. The three principal components representing the surveys were significantly predicted by measures logged during the students' interaction with ARIES.

Table II: Pearson Correlations

<b>Overall Measures</b>	<b>Positive</b>	<b>Negative</b>	<b>Perceived Learning</b>
Match score	0.181*	-0.289**	0.012
WordCount	0.231**	0.057	0.057
Gender (Male =1, Female=0)	0.031	-0.151*	0.105
Time spent reading book per chapter	0.109*	-0.161*	0.250**
Match score: Definition sub-section	0.092*	-0.240**	0.082
Match score: Importance sub-section	0.187*	-0.188*	-0.005
Match score: Example sub-section	0.113*	-0.189*	0.001
Wordcount: Definition sub-section	0.231**	0.057	0.06
Wordcount: Importance sub-section	0.203**	0.066	0.059
Wordcount: Example sub-section	0.203**	0.058	0.045
Time spent Definition sub-section	0.061	0.130*	0.082
Time spent Importance sub-section	0.001	0.072	0.085
Time spent Example sub-section	0.043	0.018	0.007
Type of Trialog Definition sub-section	0.125*	-0.081	-0.027
Type of Trialog Importance sub-section	0.070	-0.056	0.018
Type of Trialog Example sub-section	0.082	-0.138	-0.045
MC Performance per Definition sub-section	-0.087*	0.104	-0.051
MC Performance per Importance sub-section	0.033	-0.085	0.056
MC Performance per Example sub-section	0.055	-0.198*	-0.031

\*= statistical significance at  $p < .05$

\*\* = statistically meaningful in our analyses  $r > .2$

### 5.3 Mixed Effect Models

A series of linear mixed-effects models [Pinheiro and Bates 2000] were conducted in order to investigate the effects of match score (quality), number of words (quantity) and reading times from the chapters of ARIES on the three principal components associated with the survey ratings. They included the amount of words used by the student within a chapter (quantity), the cumulative match score of the student's response to the initial question in all three conversations included in one chapter (quality), reading times and the chapters themselves. The chapters themselves were treated as a fixed factor because we wanted to know if specific chapters may affect survey ratings more than others. We were not attempting to generalize across all chapters of research methodology, but rather inspect the impact of the variable topics within ARIES. Conversely, we were attempting to generalize across participants. Therefore, participant was always held as a random factor according to a standard R function for random effects (lme4 package in R, lmer function).

Random effects are specified with the assumption that we are selecting these levels of these factors from a population (not a fixed set). By using a random effect we expect our results to be more generalizable, since effects will be less likely to be detected that are merely random differences due to selection. Thus, the inherent differences of the participants can be partialled out when analyzing the fixed effects of interest. The dependent measures included the three principal components (i.e. positive valence, negative valence and perceived learning, respectively). The following analyses show the different conversational measures uniquely predict the positive and negative valence principal components. However, neither the match score, number of words, nor reading times were found to significantly predict perceived learning.

An advantage of using mixed models is the ability for researchers to compare models. The goal is not always to find the best fit, but rather to instruct the researchers as to the exact nature of the effects in the additive model. For this reason, a series of models for each criterion variable will be presented in order to inform the reader about the additive nature of the effects. Practically speaking, this allows for us to see which predictors increase reports of positive or negative affective valences. In order to accomplish this goal, an estimation of variance accounted for (represented by  $R^2$ ) was computed for each model using a reml type correction [Stevens 2007] that was inserted in the R programming code. The reml correction was used in hopes of gaining an estimate of variance accounted for within a mixed model.

We began our analyses by measuring the effects of reading times on the three principal components of perceived learning, positive valence and negative valence. Reading times proved not to predict perceived learning as the variable did not contribute any variance nor was it significantly different from the null model distribution of student scores ( $R^2 < .001, p > .25$ ). The predictor of reading times was not significant in accounting for negative valence or positive valence when compared to the best fit models for each principal component (i.e. word count and chapter for positive valence and match score and chapter for negative valence). Therefore, we continued our investigation concentrating on the unique contributions of the conversational measures of *quality* and *quantity* of words, represented by match score and word count generated by the student.

We conducted a series of analyses by comparing models that had different sets of predictor variables. The most rigorous test of any given predictor variable would be the incremental impact of that variable after removing

the contributions of student variability, chapter variability, and other predictors. More specifically, the goal of these analyses is to determine whether the conversational predictors of quality and quantity of words generated can uniquely predict the principal components associated with the ratings (positive valence, negative valence, perceived learning).

5.3.1 *Positive Valence.* A comparative series of models revealed that match score was not significantly different from the distribution of student scores (null model) and chapter alone, contributing near zero percent of the variance ( $R^2 < .001$ ) beyond student and chapter. However, the addition to these models of number of words yielded a significant unique contribution of 2% of the variance as seen in Table III. So number of words explained an increment in variance accounted for over and above the other factors, even when words are at a disadvantage of being stripped of any commonality of variance shared with students, chapters, and match scores. This is the strongest possible test for the claim that students who generate more words tend to have self-reports of positive affect.

Table III. Word Count Predicts Positive Valence

Model	DF	Obs	LL	AIC	R <sup>2</sup>	Pr(Chisq)
Null	3	523	-584.70	1175.4	.378	
Chapter	23	523	-558.67	1163.3	.418	5.927e-05**
Chapter+ Match	24	523	-558.44	1164.9	.418	0.471
Chapter+Words+Match	25	523	-555.37	1150.8	.436	.0003745***

5.3.2. *Negative Valence.* A comparison series of models revealed that the quantity of words generated during a chapter differed significantly from the distribution of student scores (null model) and chapter alone but only contributed less than 1% of the variance ( $R^2 = .009, p < .05$ ). Alternately, the addition to these models of match score (or quality of input) added nearly 2% of the variance and was significant at the level of ( $p < .001$ ) from both the null model and the model including chapter plus the distribution of student scores as seen in Table IV. In these analyses, quality of words explained an increase in negative affect variance above and beyond the other factors.

Table IV: Match Score Predicts Negative Valence

Model	DF	Obs	LL	AIC	R <sup>2</sup>	Pr(Chisq)
Null	3	523	-593.14	1192.3	.394	
Chapter	23	523	-557.33	1160.7	.477	9.879e-08***
Chapter+ Words	24	523	-552.52	1153.0	.476	0.0019182**
Chapter+Words+Match	25	523	-546.05	1142.1	.490	0.0003217***

5.2.3 *Perceived Learning.* A comparative series of models revealed that none of the predictors other than chapter significantly predicted perceived learning. The number of words, match score, chapter, and distribution scores of students, did not differ significantly from a model including the null and chapter ( $p = .482$ ). However, chapter did significantly differ from the null model and predicted 6.7% of the variance as seen in Table V.

Table V: Chapter Predicts Perceived Learning

Model	DF	Obs	LL	AIC	R <sup>2</sup>	Pr(Chisq)
Null	3	523	-636.63	1279.3	.382	
Chapter	23	523	-609.10	1264.2	.444	4.017e-05 ***
Chapter+ Words+Match	25	523	-608.37	1266.7	.445	0.482

#### 5.4 Main Effects and Cross-Validation

After considering all of the components of the analytical approaches reported above, the best fit models for all three outcome measures were extracted, taking significance and model comparison into account. After completing these analyses, the results were cross-validated using a 10-fold cross-validation procedure. Linear models with only fixed effects were used for the cross-validation because methods are not yet available for predicting a test set with a random factor. In each linear 10-fold analysis, the participants were randomly chosen to go in each of the ten folds. Then each fold was used as a test set and the others used to train the model. The resulting correlations were an average of the results for each of the folds. An  $R^2$  is reported although the original output was computed with a Pearson  $r$ . Before giving the results from the cross-validation, the main effects for the best fit mixed models will be reported for each of the principal components: positive valence, negative valence, and perceived learning.

As the reader may recall, a mixed model with the two fixed factors of words generated and chapter with participant held as a random factor predicted the positive valence factor. This model yielded a significant main effect for number of generated words ( $F(1, 522) = 6.96, p < .001$ ) and a significant main effect for chapter ( $F(20, 503) = 3.019, p < .001$ ). The number of words generated by the student has a positive relationship with the ratings on positive affect ( $t = 3.401, p < .001$ ), suggesting that the more words generated by the student correlates with positively-valenced perceptions of the learning experience. Two variations of this model were cross-validated using a 10-fold cross-validation procedure. First, a model with just number of words as a predictor of positive valence was cross-validated yielding a training set accounting for 2% of the variance ( $R^2 = .02$ ) and a test set for 2% of the variance ( $R^2 = .02$ ). Next, the model of chapter and number of words predicting positive affect resulted in a training set accounting for 5% of the variance ( $R^2 = .05$ ) and a test set accounting for 3% of the variance ( $R^2 = .03$ ). These results suggest that solely the number of words generated can predict the positively-valenced principal component as well as a multi-factor model of chapter and words.

A second mixed model with the two fixed factors of quality of words and chapter with participant held as a random factor predicted the negative valence principal component. The model yielded a significant main effect for chapter ( $F(20, 502) = 2.28, p < .001$ ) and match score corresponding to the initial student answer ( $F(1, 522) = 44.982, p < .001$ ). The quality of the words has a negative relationship with negative affect ( $t = -3.88, p < .001$ ), suggesting that lower match scores predict more negatively-valenced perceptions of the learning experience. Several variations of this model were cross-validated using a 10-fold cross-validation procedure. First, a model including only match score as a predictor of negative valence was cross-validated yielding a training set accounting for 8% of the variance ( $R^2 = .08$ ) and a test set accounting for 7% of the variance ( $R^2 = .07$ ). Next, the entire model of match score and chapter predicting negative affect produced a training set accounting for 12% of the variance ( $R^2 = .12$ ) and a test set accounting for 10% of the variance ( $R^2 = .10$ ). As the reader may recall, word count appeared to

be statistically significant during model comparison. However, it accounted for less than 1% of the variance. Therefore, a third model including words as the predictor and the negative principal component as the criterion variable was cross-validated using the ten-fold cross-validation procedure. This model was limited as the training set accounted for 1.7% of the variance ( $R^2 = .017$ ) and the test set accounted for 3% of the variance ( $R^2 = .029$ ). Upon further analysis, the relationship between negative affect and words contributed was heavily skewed, thus suggesting that a few outliers may be contributing to the results yielding a test set that far outweighs the training set. If words do account for any of the variance of negative affect, the amount is extremely low and likely a result of Type I error. With these results, we can conclude that *quality* rather than *quantity* of words and chapter are most predictive of negative affect.

None of the predictor variables of match score, reading times, or quantity of words created a predictive model of perceived learning. Yet, chapter had a significant main effect ( $F(20,502) = 2.797, p < .01$ ). However, the predictive model cannot be substantiated upon cross-validation with a linear model including only chapter as the predictor of learning. In fact, the training set produced a positive correlation ( $r = .23$ ) and the test set produced a negative correlation ( $r = -.08$ ). These results suggest that we are unable to predict perceived learning with our current set of measures logged within ARIES.

## 6. LIMITATIONS

It is important to acknowledge some salient limitations of the present study. As with all initial field trials, the study was conducted on a small sample of courses (only 2) and teachers (1), so there is a question of the generality of the findings over teachers, classrooms, and student populations. Another limitation was dropout rate of students because of non-compliant students or life circumstances that yielded missing data, reducing the original observations from 967 to 523. Linear Mixed Models were used in an attempt to adjust for these two issues, but there is the ubiquitous worry of selection biases in the face of noticeable attrition. But these standard limitations occur in any field trial and the present study does not noticeably stand out in these challenges.

Linear Mixed Effects models are particularly useful when a dataset is missing observations or it is reasonable to assume variation across participants. The computational method allows one to analyze the data at the grain-size of individual observations, thereby increasing the power of the regression models. Mixed models also allow for one to enter random factors, which generate a separate error term for each participant, the random factor in this analysis. This extra constraint decreases the probability of Type I error and increases the ability of researchers to generalize results across participants. However, the concomitant limitation is that some of the assumptions of these statistical models may be called into question and give reason to pause about some of the conclusions.

The study is further limited in the scope of variables tested. Affective measures were available for one module of ARIES, which is perfectly fine, but the researchers restricted their vision to the quantitative measures logged within the system rather than including additional measures that analyzed the textual input. Indeed, the only measure motivated by computational linguistics algorithms was the semantic match score. Many more analyses could be conducted on the text content, such as the measures derived from CohMetrix [Graesser and McNamara

2011]. Adding further computational linguistics measures on string variables are reserved for future studies, but run the risk of increasing the probability of Type I errors.

The data was collected in the context of an ecologically valid environment rather than an experimental setting with a control group. There was no comparison to a control group with random assignment of participants to condition. Therefore causal claims cannot be made from this study with respect to the impact of cognitive-discourse variables on the affect and metacognitive impressions. Nevertheless, prediction in the correlational sense is an important first step before one invests the resources into randomized control trials, as nearly all researchers agree. We are safe in making the claims that quantity of words generated by students predicts positive affect and the quality of words predicts negative affect. This is a solid claim, although the effect sizes of these trends are modest. The modest effect sizes are important to acknowledge, but it is also important to recognize that many theoretically important variables in psychology have a modest impact in dependent variables and also that important predictors in the laboratory are often subtle when one moves to field testing.

There is little doubt that video games have a strong influence on player's emotional reactions. As creators of serious games, we are attempting to blend these game-like aspects with pedagogical principles such as tutorial dialog. Indeed it is a lofty goal to attempt to blend fun with learning, especially because these two constructs are so divergent in nature [Graesser et al. 2009; McNamara et al. 2012]. Furthermore, discovering which features might best predict affect in such a highly intertwined learning environment is certainly a challenge. We have found two cognitive-discourse predictors related to the pedagogical conversations as predictors of affect in this study: quality and quantity of student contributions. Follow up studies with larger data sets will be necessary to understand the complex nature of emotions and learning within a serious game.

## 7. CONCLUSIONS

The results suggest specific cognitive-discourse features were unique predictors of positive and negative affective valences. Specifically, quantity (as measured by number of words) predicted positive valences, whereas quality of the human contribution had the most predictive power for negative valences. Additionally, the specific chapter or topic of research methodology about which the student engaged in tutorial dialog contributes to both positively and negatively-valenced impressions. However, none of these measures consistently predicted the students' perception of learning, a metacognitive criterion variable.

The relations between emotions and learning have been investigated during the last decade in interdisciplinary studies that intersect computer science and psychology [Baker et al. 2010; Conati and MacClaren 2010; D'Mello and Graesser 2012; Graesser and D'Mello 2012]. We predicted that differences may exist between student ratings of game-like versus pedagogical features of the game. For example, students may have had both high and low ratings of positive valence at the same time (i.e. loved the storyline but frustrated by the trialogs). However, the principal component analysis did not group affective responses into these two categories, but rather it did group them into positive versus negative-valenced affective states, each of which included both game-like and pedagogical features. The analysis was performed on ratings that were measured after interaction with each of the 21 chapters. Therefore, the same student could have reported a negative affective state after interaction with one chapter and a

positive state after another. The attributions of these ratings are not based on the traits of the students but rather transient states occurring across interaction with ARIES. Therefore, predictions based on cognitive-discourse features that occur during tutorial conversations were predicting dynamic affective states directed at both the entertaining and pedagogical aspects of the game.

One explanation for the current findings is that the experience of flow occurred during interactions with some chapters because there was an appropriate pairing of skills and tasks. That is, the data were compatible with the claim that the student maintained a heightened level of engagement rather than being bored, confused, or frustrated. The occurrence of this emotion has also been found also in AutoTutor when students are generating information and the feedback is positive [D'Mello and Graesser 2010;2012]. More specifically, positive affective reports could be indicative of flow when the quantity of contributions from students is both high and quickly generated. Students who were not experiencing this positive state may simply want to complete the task with lower output, with less motivation, and possibly the affective states of frustration or boredom. The students experiencing these negative emotions may be discouraged or not care about contributing to the conversation in light of the comparatively low output, the negative feedback, or low meshing with a deep understanding of the material that is reflected in accurate discriminations. These students apparently experience negative affect by virtue of their inability to articulate the ideal answer and the lack of positive feedback from the pedagogical agent. The claims on positive and negative affect of course require replication in other learning environments that will help us differentiate the relative roles of generation, feedback, and understanding.

The predictions of positive and negative affect may also be explained by the students' beliefs, specifically the student's self-schema of self-efficacy. These ingrained beliefs are known to influence the amount of effort the student expends towards any given task [Dweck 2002] and students with high self-efficacy are not likely to be affected by immediate contradictory feedback [Cervone and Palmer 1990; Lawrence 1988]. In the present study, perhaps students experiencing high levels of self-efficacy were exerting more effort as indicated by the generation of more words within the tutorial context. The discriminatory match scores that determine the tutorial agent's feedback may not sway these higher efficacy students. That is, feedback does not influence self-efficacy levels when the student has high self-confidence [Bandura 1997]. However, students experiencing lower self-efficacy levels may feel less hope in their ability to understand the material when negative feedback is given. These are circumstances that may account for the negative affective reports

The formulated self-belief is not necessarily a trait of a student but rather a dynamic state experienced by the student throughout the learning experience [Bandura 1997]. Students may feel a high level of self-efficacy towards learning one topic but feel incapable of successfully learning another topic. This personal monitoring of performance within and across the topics may play a role in self-efficacy levels towards mastering the current topic. The students may feel that they could have performed better on the most recent topic and that the learning progression has slowed, thus leaving the students' feeling like further effort may not yield any gains. This explains why students may vacillate between different cognitive-discourse behaviors and emotions across different topics. In order to test this hypothesis in future studies, the monitoring of task choices may be enlightening. Students with

higher levels of self-efficacy are more likely to choose difficult tasks whereas less assured students will be less likely to take academic risks [D’Mello et al. 2012].

Another explanation for these findings is the *broaden and build theory* [Frederickson 2001] which predicts different cognitive perspectives as correlating with positive versus negative-valenced emotions. Positive emotions are correlated with more a creative thought process and global view [Clare and Huntsinger 2007, Frederickson and Branigan 2005, Isen 2008] whereas negative emotions are correlated with a localized, methodical approach [Barth and Funke 2010; Schwarz and Skumik 2003]. Higher word generation, regardless of positive or negative feedback, could possibly be a reflection of a student experiencing heightened levels of creativity and corresponding positive emotions. Conversely, students in a negative-valenced affective state may approach the topic methodically and therefore have a greater dependence on the feedback of the pedagogical agent. Students experiencing this state are hyper-focused on circumventing the immediate obstacle

In future studies, we may be able to differentiate between these three hypotheses by altering the task the participants perform. In the current investigation, the task involved articulation of didactic knowledge about the texts that were read in chapters and pedagogical tutorial agents to facilitate the articulation of this knowledge. This may be a task conducive to flow because production of verbal output was encouraged whereas making fine discriminations about deep knowledge was not emphasized. One could set up alternate tasks in which theoretical positions other than flow (e.g., self-schema, broaden and build) could be more sensitively tested. Perhaps a more difficult and discriminating task with shorter interactions could provide such a contrast in task. Specifically, students could interact with the Case Study module of ARIES in which one must distinguish between flaws in multiple research cases, each taking approximately 15 minutes. In this module, students have natural language tutorial conversations with animated agents in order to discover these flaws (e.g. correlation versus causation). During interaction, students are given immediate and frequent negative feedback after each research case, typically resulting in decreased self-efficacy. If the difficulty of the cases was manipulated, then perhaps self-efficacy would reign supreme rather than the mere generation on verbal content in a comparatively swift manner. The task ascribed in the Case Study module requires fine-grained analytical processing of the problem rather than a creative production of information. According to the broaden and build theory, students participating in this task should experience more negative emotions. If students experience high match scores on generated input along with negative affect during this task, then perhaps the theory could explain the current findings of quality rather than quantity of words predicting negative affective states.

Several measures were not found to be predictive of affective states, notably performance on multiple-choice tests, the type of trialog, latency measures, and gender. In view of all of the explanations explaining the match score as a predictor of negative affect, one may see how multiple-choice questions could have also been a significant predictor. However, a high degree of multicollinearity existed between the variables of match score and multiple-choice performance, so one possible conclusion is that the match score was simply the better predictor between the two. Additionally, the type of trialog is correlated with the answers to the multiple-choice questions because it is algorithmically tied to the performance on these questions. Therefore, a similar collinearity explanation may account for these findings as well. The latency measures of time spent on task did not correlate with either

emotion. It is possible that students could spend more time on one topic because they were in a state of flow. Conversely, a student may spend more time during tutorial dialog because the topic is too difficult and the student is unable to provide the correct answer. Therefore, the time spent on each topic could be attributed to an experience of positive or negative affect, thus making them predictive of neither. Gender was also found to be a non-correlate of the affective states. However, the trend shows less negative and more positive ratings for men than women. No strong theoretical claims are made about the trends of gender but they are reported due to the current interest in the topic. The current study did not show relations between affective states and the measures of performance on multiple-choice tests, type of dialog, the latency measure, and gender, nor were there significant relations with the metacognitive principal component. However, it should be kept in mind that other research has found some of these factors to correlate with emotions [D'Mello and Graesser 2012], so the findings of the current study cannot conclusively rule out any existing relationships.

The current study found two cognitive-discourse features which predict affective valences, but none of the logged measures affected the third principal component of perceived learning characterized by judgments of learning and perceived chapter difficulty. It is not surprising that perceived learning does not appear to have a significant predictor. The metacognition literature shows that students often lack the ability to assess their own learning [Dunlosky and Lipko 2007; Graesser et al. 2009; Maki 1998]. Thus, students may also be unable to assess performance accurately during the learning experience. With such a deficit of perception, it is no wonder that perceived learning can be differentiated effectively from positive and negative emotions. Thus, these specific conversational predictors of affective states are unable to predict perceived learning. The time spent reading the E-book was the most likely predictor of perceived learning. However, the relationship between reading a text and learning is complex including many factors such as prior-knowledge [McNamara 2001; McNamara and Kintsch 1996], textual features [Graesser and McNamara 2001] and length of text [Mayer and Moreno 2003; Sweller and Chandler 1994]. There may simply be too many confounding variables in order to discover relationships between reading an E-book and self-reports of metacognition and affective states during learning within a serious game.

Overall, the assumption that performance features are predictive of affective states has been substantiated. We were unable to predict metacognitive responses but there are good reasons why this would be expected. Regardless, the game aspects of ARIES did not over-ride the predictive nature of discourse features found within the tutorial conversations. These findings support the proposition that the interactions within ARIES may be indicative of affective states over and above the inherent game-like nature of the system.

## REFERENCES

ARROYO, I., WOOLF, B., COOPER, D., BURLESON, W., MULDER, K., AND CHRISTOPHERSON, R. 2009. Emotion sensors go to school. In *Proceedings of the 14th International Conference on Artificial Intelligence In Education*, Brighton, UK, July 2009, V. DIMITROVA, R. MIZOGUCHI, B. DU BOULAY & A. GRAESSER, Eds. IOS Press, Amsterdam, Netherlands, 17-24.

- ATKINSON, R. K. 2002. Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology* 94, 416-427.
- BAKER, R. S. J., CORBETT, A. T., KOEDINGER, K. R., EVENSON, S., ROLL, I., WAGNER, A. Z., NAIM, M., RASPAT, J., BAKER, D. J., AND BECK, J. E. 2006. Adapting to when students game an intelligent tutoring system. In *Proceedings of Intelligent Tutoring Systems 8th International Conference*, M. IKEDA, K. D. ASHLEY, AND T.W. CHAN, Eds. ITS 2011, May 2011, Springer-Verlag, Berlin, Germany, 392-401.
- BAKER, R. S. J., AND DE CARVALHO, A. M. J. A. 2008. Labeling student behavior faster and more precisely with text replays. In *Proceedings of the 1st International Conference on Educational Data Mining*, Montreal, Canada, June 2008, R. S. J. BAKER, T. BARNES, AND J. E. BECK Eds. 38-47.
- BAKER, R.S.J., D'MELLO, S.K., RODRIGO, M.T., AND 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.
- BANDURA, A. 1997. *Self-efficacy: The Exercise of Control*. Freeman, New York, NY.
- BARRETT, L. F. 2007. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review* 10, 20.
- BARTH, C.M. AND FUNKE, J. 2010. Negative affective environments improve complex solving performance. *Cognition and Emotion*, 24, 1259-1268.
- BAYLOR, A. L. AND KIM, Y. 2005. Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education* 15, 95-115.
- BISWAS, G., SCHWARTZ, D. L., LEELAWONG, K., VYE, N., AND TAG, V. 2005. Learning by teaching: A new paradigm for educational software. *Applied Artificial Intelligence* 19, 363-392.
- BROWN, A. L., ELLERY, S., AND CAMPIONE, J. C. 1998. Creating zones of proximal development electronically. In *Thinking practices in mathematics and science learning*, J.G. GREENO AND S. V. GOLDMAN, Eds. Lawrence Erlbaum, Mahway, NJ, 341-367.
- BRUNER, E.M 1986. Ethnography as narrative. In *The Anthropology of Experience* V.W. TURNER AND E.M BRUNER, Eds. University of Illinois Press, Urbana, OH 139-155.
- CAI, Z., GRAESSER, A.C., FORSYTH, C., BURKETT, C., MILLIS, K., WALLACE, P., HALPERN, D. & BUTLER, . 2011. Dialog in ARIES: User input assessment in an intelligent tutoring system. In *Proceedings of the 3<sup>rd</sup> IEEE International Conference on Intelligent Computing and Intelligent Systems*, Guangzhou, China, November 2011, W. CHEN & S. LI, Eds. Guangzhou: IEEE Press, 429-433.
- CALVO, R. A., AND D'MELLO, S. K. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing* 1, 18-37.
- CERVONE, D. AND PALMER, B.W. 1990. Anchoring biases and the perseverance of self-efficacy beliefs. *Cognitive Therapy and Research* 14, 401-416.
- CLORE, G.L. AND HUNTSINGER, J.R. 2007. How emotions inform judgment and regulate thought. *Trends in Cognitive Science* 11, 393-399.

- CONATI, C., CHABBAL, R. & MACLAREN, H. 2003. A study on using biometric sensors for detecting user emotions in educational games. In *Proceedings of the Workshop Assessing and Adapting to User Attitude and Affects: Why, When and How?*.
- CONATI, C. AND 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'MELLO, S., WITHERSPOON, A., AND GRAESSER, A.C 2008. Emote aloud during learning with AutoTutor: Applying the facial action coding system to cognitive-affective states during learning. *Cognition and Emotion 22*, 777-788.
- CRAIG, S., GRAESSER, A., SULLINS, J., AND GHOLSON, J. 2004. Affect and learning: An exploratory look into the role of affect in learning. *Journal of Educational Media 29*, 241-250.
- CSIKSZENTMIHALYI, M. 1990. *Flow: The Psychology of Optimal Experience*. Harper-Row, New York, NY.
- D'MELLO, S.K., CRAIG, S.D., AND GRAESSER, A.C. 2009. Multi-method assessment of affective experience and expression during deep learning. *International Journal of Learning Technology 4*, 165-187.
- D'MELLO, S. AND GRAESSER, A.C. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-adapted Interaction 20*, 147-187.
- D'MELLO, S. AND GRAESSER, A.C. 2012. Emotions during learning with AutoTutor. IN P.J. DURLACH AND A. LESGOLD, *Adaptive technologies for training and education*. Cambridge University Press, England, UK.
- D'MELLO, S., AND GRAESSER, A. in press. Dynamics of affective states during complex learning. *Learning and Instruction*.
- D'MELLO, S., GRAESSER, A.C. AND STRAIN, A.C. in press. Emotions in advance learning technologies. In *Handbook of Emotions and Education* R. PEKRUN AND LINNENBRING-GARCIA, Taylor and Francis Publishers, New York, NY.
- DUNLOSKY, J., AND LIPKO, A. 2007. Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science 16*, 228-232.
- DWECK, C. S. 2002. Beliefs that make smart people dumb. In *Why smart people do stupid things*. R. J. Sternberg Ed. New Haven, CT: Yale University Press.
- FREDERICKSON, B. L. 2001. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist 56*, 218-226.
- FREDERICKSON, B.L. AND BRANIGAN, C. 2005. Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition and Emotion 19*, 313-332.
- GEE, J.P. 2003. *What Video Games Teach Us about Language and Literacy*. Palgrave/Macmillan, New York, NY.
- GRAESSER, A. C., CHIPMAN, P., LEEMING, F., & BIEDENBACH, S. 2009. Deep learning and emotion in serious games. In *Serious games: Mechanisms and effects*, U. RITTERFELD, M. CODY, AND P. VORDERER, Eds. Taylor & Francis, Routledge: New York and London, 81-100.
- D'MELLO, S. AND GRAESSER, A.C. in press. Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies*.

- GRAESSER, A.C. AND D'MELLO, S. 2012. Emotions during the learning of difficult material. In. B. ROSS Eds., *The Psychology of Learning and Motivation 57*, Elsevier, Amsterdam.
- GRAESSER, A. C., D'MELLO, S. K., CRAIG, S. D., WITHERSPOON, A., SULLINS, J., MCDANIEL, AND 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., AND PERSON, N. K. 2009. Metaknowledge in tutoring. In *Handbook of metacognition in education* .D. HACKER, J. DONLOSKY, & A. C. GRAESSER, Eds. Taylor & Francis, New York, NY, 361-382.
- GRAESSER, A. C., JEON, M., AND DUFTY, D. 2008. Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes 45*, 298–322.
- GRAESSER, A.C., LU, S., JACKSON, G.T., MITCHELL, H., VENTURA, M., OLNEY, A., AND LOUWERSE, M.M. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers, 36*, 180-193.
- GRAESSER, A.C. AND MCNAMARA, D.S. 2011. Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3*, 371-398.
- GRAESSER, A.C., MCNAMARA, D., AND LOUWERSE, M. in press. Methods of automated text analysis. In: MICHAEL KAMIL, DAVID PEARSON, ELIZABETH MOJE, AND PETER EFFLERBACH, Eds. *The Handbook of Reading Research*, Routledge/Erlbaum, Mahwah, NJ.
- GRAESSER, A. C. AND OTTATI, V. 1996. Why stories? Some evidence, questions, and challenges. In *Knowledge and Memory: The Real Story*, R. S. WYER, Ed. Hillsdale, NJ: Erlbaum.
- GRAESSER, A. C., SINGER, M., & TRABASSO, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-395.
- GRATCH, J. AND MARSELLA, S. 2001. Tears and fears: Modeling emotions and emotional behaviors in synthetic agents. In *Proceedings of the 5th International Conference on Autonomous Agents*, Montreal, Canada, June 2001, E. ANDRE, S. SEN , C. FRASSON, J.P. MULLER, Eds. ACM, New York, NY, 278-285.
- HUANG, W.D, AND JOHNSON, T. 2008. Instructional game design using cognitive load theory. In *Handbook of Research on Effective Electronic Gaming in Education*, R. FERDIG, Ed. Information Science Reference, Hershey, PA, 1144-1164.
- HUANG, W.D. AND TETTEGAH, S. 2010. Cognitive load and empathy in serious games: a conceptual framework. In *Gaming and Cognition: Theories and Practice from the Learning Sciences*, R. VAN ECK Ed., Information Science Reference, Hershey, PA, 137-151.
- ISEN, A. 2008. Some ways in which positive affect influences decision making and problem solving. In *Handbook of emotions 3<sup>rd</sup> edition*, M. LEWIS, J. HAVILAND-JONES AND L. BARRETT, Eds. Guilford, New York, NY, 548-573.
- ISENHOWER, R.W., FRANK, T. D., KAY, B. A., AND CARELLO, C. 2010. A dynamical systems approach to emotional experience: Relating affective and event valence. In *Proceedings of the 20th Annual New England Sequencing and Timing Conference*, March 2010, New Haven, CT.

- JURAFSKY, D., AND MARTIN, J. 2008. *Speech and language processing*. Prentice Hall, Englewood, NJ.
- KAPOOR, A., BURLESON, W., AND PICARD, R. 2007. Automatic prediction of frustration. *International Journal of Human Computer Studies* 65, 724-736.
- LANDAUER, T., MCNAMARA, D. S., DENNIS, S., AND KINTSCH, W. 2007. *Handbook of Latent Semantic Analysis.*, Erlbaum, Mahwah, NJ.
- LAWRENCE, C. P. 1988. *The perseverance of discredited judgments of self-efficacy: Possible cognitive mediators*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.
- LAZARUS, R. 2000. The cognition-emotion debate: A bit of history. In *Handbook of Emotions 2nd edition*, M. LEWIS AND J. HAVILAND-JONES, Eds. Guilford Press, New York, NY, 1-20.
- LITMAN, D.J., AND 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 *Metacognition in educational theory and practice* D. J. HACKER, J. DUNLOSKY, AND A. C. GRAESSER, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 117-144.
- MALONE, T. W. AND LEPPER, M. R. 1987. Making learning fun: A taxonomy of intrinsic motivations for learning. In *Aptitude, Learning and Instruction: III. Conative and affective process analyses* R. E. SNOW AND M.J. FARR, Eds. Erlbaum, Hillsdale, NJ, 223-253.
- MANDLER, G. 1999. Emotion. In B.M BLY AND D.E RUMELHART (Eds.), *Cognitive Science Handbook of Perception and Cognition 2nd edition*. San Diego, CA: Academic Press.
- MAYER, R.E. in press. Narrative games for learning: Testing the discovery and narrative hypotheses. *Journal of Educational Psychology*.
- MAYER, R. E. AND ALEXANDER, P. A. Eds. 2011. *Handbook of Research on Learning and Instruction*. Routledge, New York, NY.
- MAYER, R. E. AND MORENO, R. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38, 43-52.
- MCNAMARA, D.S. 2001. Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, 55, 51-62.
- MCNAMARA, D. S. AND KINTSCH, W. 1996. Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, 22, 247-287.
- MCNAMARA, D.S., O'REILLY, T., ROWE, M. BOONTHUM, C. AND LEVINSTEIN, I.B. 2007. iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In *Reading comprehension strategies: Theories, interventions, and technologies*, D. S. MCNAMARA Ed. Erlbaum, Mahwah, NJ, 397–421.
- MCNAMARA, D. S., RAINE, R., ROSCOE, R., CROSSLEY, S., JACKSON, G. T., DAI, J., CAI, Z., RENNER, A., BRANDON, R., WESTON, J., DEMPSEY, K., LAM, D., SULLIVAN, S., KIM, L., RUS, V., FLOYD, R., MCCARTHY, P. M., & GRAESSER, A. C. 2012. The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In *Applied natural language processing: Identification, investigation, and resolution* P. M. MCCARTHY AND C. BOONTHUM-DENECKE Eds. IGI Global, Hershey, PA, 298-311.

- MCQUIGGAN,S., ROBINSON, J. AND LESTER,J. 2010. Affective transitions in narrative-centered learning environments. *Educational Technology & Society*, 13, 40-53.
- MCQUIGGAN,S., ROWE,J., LEE, S., and LESTER,J. 2008. Story-Based Learning: The Impact of Narrative on Learning Experiences and Outcomes. In *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems (ITS-08)*, Montreal, Canada, June 2008, B.P. WOOLF, E. AIMEUR, R. NKAMBOU, S.P. LAJOIE Eds. Lecture Notes in Computer Science, Springer Verlag Heidelberg, Germany, 530-539.
- MILLIS, K, FORSYTH, C., BUTLER, H., WALLACE, P., GRAESSER, A.,AND HALPERN, D. 2011. Operation ARIES! A serious game for teaching scientific inquiry. In *Serious games and Edutainment Applications*, M. MA, A. OIKONOMOU & J. LAKHMI, Eds. Springer-Verlag, London, UK, 169-196.
- MORENO, R. AND MAYER, R. E. 2004. Personalized messages that promote science learning in virtual environments. *Journal of Educational Psychology* 96, 165-173.
- NICAUD, J.F., BOUHINEAU, D., AND CHAACHOUA, H. 2004. Mixing microworld and CAS features in building computer systems that help students learn algebra. *International Journal of Computers for Mathematical Learning* 9, 169-211.
- O'NEIL, H. F. AND PEREZ, R. S. 2008. *Computer games and team and individual learning*. Elsevier, Oxford, UK.
- PEKRUN, R. 2006. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review* 18, 315-341.
- ORTONY, A., CLORE, G., AND COLLINS, A. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press, New York, NY.
- PEKRUN, R., GOETZ, T., DANIELS, L., STUPNISKY, R. H., & RAYMOND, P. 2010. Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, 102, 531-549.
- PEKRUN, R., MAIER, M.A., AND ELLIOT, A.J. 2009. Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of Educational Psychology* 101, 115-135.
- PINHEIRO, J.C. AND BATES, D.M. 2000. Mixed-Effects models in S and S-PLUS. *Statistics and Computing Series*, Springer-Verlag, New York, NY.
- RATAN,R. AND RITTERFELD, U. 2009. Classifying serious games. In *Serious games: Mechanisms and effects*. U. RITTERFELD, M. CODY, and P. VORDERER, Eds. Routledge, New York, NY, 10-24.
- RITTERFELD, U., SHEN, C., WANG, H., NOCERA, L., AND WONG, W. L. (2009). Multimodality and interactivity: connecting properties of serious games with educational outcomes. *CyberPsychology and Behavior*, 12, 691-698.
- ROBINSON, D. H., LEVIN, J. R., THOMAS, G. D., PITUCH, K. A., & VAUGHN, S. R. 2007. The incidence of “causal” statements in teaching and learning research journals. *American Educational Research Journal* 44, 400–413.
- ROWE, J., SHORES, L., MOTT B., AND LESTER, J. 2011. Integrating learning, problem-solving and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education: Special Issue on Best of ITS 2010*, 21, 115-133.

- RUSSELL, J. A. 2003. Core affect and the psychological construction of emotion. *Psychological Review* 110, 145–172.
- SCHERER, K., SCHORR, A., AND JOHNSTONE, T. 2001. *Appraisal Processes in Emotion: Theory, Methods, Research*. London University Press, London, UK.
- SCHWARZ AND SKURNIK, I. chwarz, N., & Skurnik, I. (2003). Feeling and thinking: Implications for problem solving. In *The Psychology of Problem Solving*, J.E. DAVIDSON AND R.J. STERNBERG, Eds. Cambridge University Press, New York, NY, 263-290.
- SPIELBERGER, C.D. AND REHEISER, E.C. 2003. Measuring anxiety, anger, depression and curiosity as emotional states and personality traits with the STAI, STAXI and STPI. In *Comprehensive Handbook of Psychological Assessment* (Vol. 2), M.J. Hilsenroth, D. SEGAL AND M. HERSEN, Eds. John Wiley & Son, Hoboken, NJ., 70-86.
- SPIRES, H. A., TURNER, K. A., ROWE, J., MOTT, B., AND LESTER, J. 2010. *Game-based literacies and learning: Towards a transactional theoretical perspective*. Paper presented at the meeting of the American Educational Research Association, Denver, CO, May 2010.
- STEVENS, H 2007. R-squared code for lmer function. <http://www.mail-archive.com/r-help@stat.math.ethz.ch/msg86633.html>.
- SWELLER, J. AND CHANDLER, P. 1994. Why some material is difficult to learn. *Cognition and Instruction*, 12, 185-233.
- VANLEHN, K., GRAESSER, A. C., JACKSON, G. T., JORDAN, P., OLNEY, A. M., AND ROSE, C. 2007. When are tutorial dialogues more effective than reading? *Cognitive Science* 31, 3-62.
- VORDERER, P. AND BRYANT, J. Eds. 2006. *Playing Video Games: Motives, Responses, and Consequences*. Lawrence Erlbaum Associates, Mahwah, NJ. IOS Press
- VYGOTSKY L. S. 1986. *Thought and Language*. MIT Press, Cambridge, MA. (Original work published 1934).
- WEINER, B. 1992. *Human motivation: Metaphors, theories and research*. SAGE Publications, Newbury Park, CA .
- WHITTON, N. 2010. *Learning with digital games: A practical guide to engaging students in higher education*. Routledge, New York, NY.
- ZIMMERMAN, B. J. AND SCHUNK, D.H. 2008. Motivation: An essential dimension of self-regulated learning. In *Motivation and Self-Regulated Learning: Theory, Research, and Applications*, D. H. SCHUNK AND B. J. ZIMMERMAN, Eds. Erlbaum, Mahwah, NJ, 1-30.

## APPENDIX A: ARIES TRAINING MODULE INTERFACE



## APPENDIX B: EXAMPLE CONVERSATION WITH ARIES

**QUESTION:** Glass: Okay, I know operational definitions are important in research, but I'm not sure why.

**HINT:** Glass: Let's just go over the importance of these things one more time. Why do we need to have operational definitions?

Human: so that they can be reliably measured and used by other researchers

**NEUTRAL FEEDBACK:** Glass: OK.

**PUMP:** Dr. Quinn: Um, can you add to that?

Human: it is important to create operational definitions because they provide a basis for measuring variables and so that they can be reliably measured and used by other researchers in other studies.

Glass: That is not clear to me.

**HINT:** Glass: I thought operational definitions had something to do with telling others about your study. What could that be?

Human: how to measure a variable

Glass: I still don't understand.

**PROMPT:** Glass: Wow, it seems like operational definitions have something to do with being able to reliably do something. What is that something?

Human: measure a variable

**NEUTRAL FEEDBACK:** Dr. Quinn: Okay.

**CORRECT PROMPT COMPLETION:** Dr. Quinn: reliably recognize variables

**PROMPT:** Glass: I'm remembering something about how operational definitions make it so that a variable means the same thing to all researchers. That means that when other researchers use the same variable, it is what?

Human: reliable

**POSITIVE FEEDBACK:** Dr. Quinn: Great!

Dr. Quinn: it is reliable

Glass: Yeah, I understand.

**SUMMARY:** Glass: Operational definitions are important to consider because they allow a particular variable to be reliably recognized, measured, and understood by all researchers.

**EXPECTATION\*:** Operational definitions are important because they allow a particular variable to be reliably recognized measured and understood by all researchers.

\*= never explicitly stated as an expectation during conversation