Operation ARIES!: A Serious Game for Teaching Scientific Inquiry

Keith Millis
Northern Illinois University

Carol Forsyth
University of Memphis

Heather Butler
Claremont Graduate University

Patty Wallace
Northern Illinois University

Arthur Graesser
University of Memphis

Diane Halpern
Claremont McKenna College

To appear in: M. Ma, A. Oikonomou, and L. Jain (Editors), *Serious Games and Edutainment Applications*. Springer-Verlag, UK.

Send all correspondence to Keith Millis, kmillis@niu.edu, Department of Psychology, Northern Illinois University, Dekalb, IL, 60115, USA.

There is a long history of science fiction novels and movies that feature aliens from other worlds conquering our planet, either overtly as in *The War of the Worlds*, or covertly as in *The Arrival*. Aliens have also infiltrated video games, starting with *Space Invaders*, and more recently with *Aliens vs. Predator*. Extraterrestrials have also appeared in educational games and related learning experiences. For example, in *Alien Games*, girls and boys create a video game within an alien theme teaching principles of outer space (Heeter, Egidio, Mishra, Winn, & Caywood, 2007). Indeed, the idea of aliens taking over Earth is hardly a new idea.

Aliens have recently made their appearance in a serious game called *Operation ARIES!* In this game, players learn how to critically evaluate research that they encounter in various media, such as the Web, TV, magazines and newspapers. ARIES is an acronym for **A**cquiring **R**esearch **I**nvestigative and **E**valuative **S**kills. The game focuses on teaching critical thinking and scientific reasoning within scientific inquiry (the "how" of science). In particular, it teaches how to critically evaluate aspects of scientific investigations (e.g., the need for control groups, adequate samples of observations, operational definitions, etc.) and how to ask appropriate questions in order to uncover problems with design or interpretation. Scientific inquiry is crucial because it comprises the necessary steps of "science as process," the steps that scientists follow in establishing and critiquing causal claims (NSES, 1996).

Scientific inquiry is a crucial aspect of being an informed citizen living in the "information age". The public is constantly being exposed to causal claims made by scientists, advertisers, coworkers, friends, and the press via a variety of media (blogs, TV, Web, print, word of mouth). Of course, some of the claims have relatively solid scientific evidence for support, whereas others do not. In some cases, the research is well executed, but the interpretation or

conclusion drawn by the press is inappropriate, as in the case of a headline that makes a causal claim ("Wine Lowers Heart Attack Risk in Women") based on a correlational design which does not support a cause-effect interpretation (http://www.indianwineacademy.com/dm_145_item_4.asp).

In other cases, a claim is unfounded because the design of the study itself is flawed. For example, in one "experiment" aired on American TV, reporters secretly recorded a carload of teenage drivers. The footage shows them carelessly driving through stop signs while laughing and joking with each other. The conclusion made by the newscaster is that teenagers are too immature to drive, and the legal age limit for awarding driving licenses should be increased. To the trained eye, however, this was a poor experiment – in fact, it was not an experiment at all. There was no comparison group of older drivers, no mention of confounds (driving errors could solely decrease with the amount of driving experience rather than the age of the driver), and there was a small sample size. Undoubtedly, this 'experiment' led to many scared parents lecturing their kids into deeper teenage angst. Unfortunately, more serious consequences than unhappy teenagers can arise from careless thinking about science. According to the U.S. National Institute of Health (NIH), around four million U.S. adults and one million U.S. children used homeopathy and other alternative medicines in 2006, despite research showing little or no effectiveness beyond placebo effects (http://nccam.nih.gov/health/homeopathy). In some instances, people suffer or die from relying on treatments that they believe to be valid despite evidence that they are not. Knowing and applying scientific inquiry skills can literally save lives.

So, how can aliens help learners acquire scientific inquiry skills? *Operation ARIES!* is an adventure game in which intelligent tutoring technology is combined with video game attributes.

In the game, alien creatures called “Fuaths” from the Aries constellation are secretly publishing flawed research in various media outlets. By extensively publishing flawed research, they hope to confuse Earth’s inhabitants about the proper use of the scientific method. By doing so, humans would not be able achieve inter-galactic space travel, which would seriously hurt their economy. They are also surreptitiously stealing Earth’s valuable resources of water, plants, and oil to help rebuild their home world Thoth. Unfortunately for humans, Fuaths have the ability to look and act human, and so catching them is not an easy task.

This is the player’s primary objective *of Operation ARIES!*: the Federal Bureau of Science (FBS) has recruited the player to become a secret agent in the battle against the Fuaths. The player’s mission is to be able to spot flawed research that would then lead the FBS to be able to find and arrest the alien authors. Its target audience includes high school seniors, college students, and members of the military and the interested public.

This chapter describes *Operation ARIES!,* with a primary focus on how the game’s design incorporates various principles of learning found in cognitive psychology and the learning sciences. The game contains three modules (or levels): Training, Case Studies, and Interrogation. In the training module, players read an eBook accompanied by multiple choice questions and tutorial conversations. In the case studies module, players apply what they learned in the training module to realistic examples of flawed research. Lastly, in the Interrogation module, players learn to ask scientists pointed questions about their research and learn how to evaluate their answers. The storyline is advanced by emails, dialogs, and videos which are interspersed among the learning activities according to a set script. It begins with the player joining the FBS as an agent-in-training and concludes with the player helping to save the world.

Learning Principles/Design Features

Below we list and describe several learning principles, design features, and gaming characteristics that have been implemented into *Operation ARIES!* Most of these are related to one another, and certainly there are others that have been implemented but are beyond the scope of this chapter (e.g., reflection, spacing effects, authentic learning, active learning). These have been compiled by researchers who have shown them to be related to increases in learning gains, engagement, interest, or motivation.

1. **Zone of proximal development.** Vygotsky's (1978) "zone of proximal development" refers to the distance between learning that occurs by an individual working alone on a problem and the learning that results when given proper instruction and guidance. When placed outside of the zone, the learning activities are too difficult for the individual, and consequently, students experience frustration or disengagement rather than learning (Rieber, 1996). Indeed, researchers and game designers have argued that optimal learning occurs when the match between the skills acquired by the learner and the requirements of the activities is neither too easy nor too hard (Van Eck, 2007). It is in this "zone" that learners experience "flow" (Csikszentmihalyi, 2002) and "cognitive disequilibrium" with moderate confusion in the face of temporary impasses during learning (Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005). Flow experiences occur frequently while playing digital games (Benyon, Turner, & Turner, 2005), whereas both flow and cognitive disequilibrium are positively correlated with learning (Graesser, D'Mello, Craig, Witherspoon, Sullins, McDaniel, & Gholson, 2008).

2. **Self-explanation**. People do not learn much when they are bored or passive. One strategy that promotes learning is self-explanation, in which the learner (reader, player) *explains* the material to one's self. A form of self-explanation occurs when the individual explains the material to another student, constituting "learning by teaching" (Biswas, Leelawong, Schwartz, & Vye, 2005). Self-explanations include identifying the causes and consequences of states and/or actions, retrieving and incorporating relevant prior knowledge, and reasoning about the information. Self-explaining increases comprehension and has been the hallmark of several learning environments (McNamara, O'Reily, Rowe, Boonthum, & Levinstein, 2007; Meyer & Wijekumar, 2007; Palinscar & Brown, 1984). Besides comprehension, self-explanation appears to increase the ability for readers to accurately judge their understanding as assessed by a later comprehension test ("metacomprehension," Griffin, Wiley, & Theide, 2008), which is notoriously poor under normal reading conditions (Maki, 1998).

3. **Feedback.** Feedback can be given in many forms, such as corrective ("correct" "incorrect") or elaborative/formative (providing hints to help the learner provide a clearer or more complete answer), and may be expressed by points, explanations, achievements, actions in the story world, and skillometers (Oxland, 2004; Shute, 2006). Informative feedback enhances learning, motivation, engagement, and self-efficacy (Anderson, Corbett, Koedinger, & Pelletier, 2006; Harackiewicz, 1979; Kulik & Kulik, 1988).

4. **Narrative, fantasy, adventure**. Games often immerse the player in a virtual fantasy world in which the player solves problems and interacts with other real or virtual characters. Often the game play is embedded in a narrative that may contain elaborate settings, characters, goals, subgoals, obstacles, and various other plot devices. Narratives

in games are often nonlinear and interactive when the player's actions determine future story states (Whitton, 2010), allowing for a high degree of replay value (Gee, 2003).

5. **Player control**. Games have various design features that allow the player to have control over the learning environment. Allowing player control can be accomplished by giving the player several options at a given time, and by providing actions that are perceived to be influential and logical to their consequences (Malone & Lepper, 1987). Customization of the display (e.g., avatars, sounds) and the ability of the player to choose levels or difficulties of play also increase player's perceptions of control (Whitton, 2010). Although adopting user control into games is seen positively by game designers, the implementation should be clear and obvious to the user (Salen & Zimmermann, 2004; Whitton, 2010).

6. **Dialogue**. People often learn by conversations and tutorial dialogs (Chi, Siler, Jeong,Yamauchi, & Hausmann, 2001; VanLehn, Graesser, Jackson, Jordon, Olney & Rose, 2007). Engaging in a conversation entails a number of processes that engender deep learning, such as generating questions and answers, retrieving information from memory, reasoning, active processing and self-explaining. Because of the tremendous challenges posed by computers understanding natural language, there are relatively few games and computerized learning environments that enable the player or learner to converse with a virtual agent or avatar in natural language.

7. **Encoding variability**. It is important that the skills and knowledge that players practice and learn in a learning environment transfer to other situations and contexts. Although achieving transfer is notoriously difficult, providing variability in examples helps the learner to discriminate relevant and irrelevant features (Bransford, Sherwood,

Hasselbring, Kinzer, & Williams, 1990), which increases transfer to novel problems (Halpern, 2002).

Below we discuss the three modules in *Operation ARIES!*, and also how the various principles relate to the module. In some cases, we present relevant research on the modules.

Module 1: Training

In this module, the player reads and is tested on various aspects of scientific inquiry. The content is provided by an eBook titled "The Big Book of Science." In many aspects, the book is conventional since much of its content is covered by research methods texts published in the social sciences. However, it is unique in a couple of ways. First, it was written by Zlotsky Amapolis, a Fuath scientist who authored the book to teach the scientific method to other Fuath operatives working on Earth. Therefore, the book is a captured alien's spy manual. Second, because it was written for the Fuath spies, it incorporates aspects of Fuath culture in elaborations and examples. For instance, the Fuaths call Human Beings "Beans" and "nose breathers" (the Fuaths do not have noses), and the concept of sample size is illustrated with Blupblops that are plants found on Thoth.

Each chapter is dedicated to one or two important concepts in scientific inquiry. There are 21 primary concepts in all, and these are listed in Table 1. We chose the topics by surveying college and university professors who teach psychology, sociology, biology, chemistry, earth science and physics classes on what they considered critical concepts for students in their field to learn.

A screen shot of the training module is presented in Figure 1. There are two animated pedagogical agents, Dr. Quinn and Glass Tealman. Dr. Quinn is the teacher whereas Glass is a

fellow student. Both of them speak and show facial expressions, and what they say is presented in a textbox so that the player can have a written record. We chose to use animated agents for several reasons. One is that they are (virtual) humans which players can relate to on a very intuitive and personal fashion. Consequently, animated agents are very engaging. A second reason is that they have been shown to increase learning and motivation in learning environments (Atkinson, 2002; Baylor & Kim, 2005). Another function is that the story line involves the agents as protagonists. Finally, some of the interactions between the agents instantiate important pedagogical roles (Baylor & Kim, 2005). As will be described below, Glass's responses serve as a model to the low knowledge player, yet Glass also serves as a teachable student for the more-knowledgeable player.

Before each chapter, Dr. Quinn and Glass hold a brief conversation, similar to an informal chat between student and teacher before a lecture. For example, early in the game Glass mentions that his new roommate was "chanting and doing Tarot cards last night. Said he was writing a paper saying that there is scientific evidence that Tarot cards can read the cosmic time space continuum. I thought it might be a hypothesis because of connecting two variables. But I also thought it sounded wacko." Dr. Quinn explains that this is most likely pseudoscience, which turns out to be the topic of the next chapter. The content of the dialog achieves two functions. One is that it advances the story line, with depictions of new events. For example, Glass's new roommates turn out to be alien spies. Another is that they introduce the topic of the chapter which is related to the dialog and the story line. This is important because after the dialog, the players are asked whether they would like to take a "challenge test." A high score on this test allows the player to skip reading the chapter. The dialog gives the student enough context of the chapter allowing the player to make an informed decision.

We should note here that *Operation ARIES!* contains different forms of internal assessments of the player's knowledge. In the Training module alone, there are multiple types of assessments. One type occurs within each chapter. These are interactive learning activities which were created by the author of "The Big Book of Science." Readers of the book are periodically given brief problems that require constructed answers (e.g., "write down the hypothesis") or selected responses (multiple choice, matching). These are included to promote reader engagement, activity and reflection with the material. As will be elaborated below, another type of assessment is more dynamic with the player being tested, given instruction, and assessed again (Yeomans, 2008).

Challenge Tests, Trialogs and AutoTutor

The player can test out of reading a chapter by doing well on a "challenge test" that evaluates the player's knowledge of the one or two key concepts addressed by the chapter. It is comprised of multiple choice test questions that assess three levels of understanding: (1) definitional, and (2) functional knowledge of the concept, and (3) identification of proper instances of the concept. From the perspective of Bloom's (1956) taxonomy, these roughly correspond to knowledge/memory, comprehension/understanding, and application, respectively. The questions are placed into two question sets: each set contains one question from each of the three levels. The first set of three questions is given to the player if he or she elects to take the challenge test, but the player will be asked to read the chapter if his or her performance is lower than a predetermined threshold. If the player's performance is above the threshold, then the second set is given. In the case where the player does not choose to take the challenge test and elects to read the chapter immediately, both sets of questions are given after the chapter is read.

Immediately after answering each of the questions in the second set, the player engages in a three-way tutorial conversation that includes the human player, Dr. Quinn, and Glass. We refer to these as 'trialogs' because there are three conversational partners. The topic of the trialogs is the material in the multiple choice questions. Consequently, the trialogs are similar to when a teacher discusses the correct answer with students in a classroom setting. There are three types of trialogs. A *standard* trialog occurs when Glass observes Dr. Quinn teach the player. A *vicarious* trialog occurs when the player watches Dr. Quinn teach Glass. A *teaching* trialog occurs when the player teaches Glass as Dr. Quinn observes. The type of trialog that occurs for a particular question is based on the level of knowledge exhibited by the player. Low, intermediate, and high knowledge triggers vicarious, standard and teaching trialogs, respectively. The level of knowledge is defined by the number of questions that the player answers correctly.

Table 2 presents examples of the different types of trialogs. The examples were drawn from a study that is summarized below. (We should note that the participants in this study did not read the eBook beforehand, and therefore their answers might be a little sparse compared to students who had read the book.) The examples are presented for illustrative purposes. The point of the trialogs is to get the player to articulate a particular idea regarding the topic of the multiple choice question. The trialogs generally have the following order of presented information: question → hint → prompt → summary. The question is the main question to be answered by one of the agents, and is compared to an "ideal answer" which is also serves as the summary. If the answer to the question is acceptable, then the summary is given and the trialog ends. Otherwise a general hint is given. The hint tries to nudge the tutee into articulating the correct answer by providing retrieval cues. If the answer to the hint is acceptable, then the summary is given. Otherwise a prompt is given. Prompts narrow down the problem space to a particular

word or phrase. After the prompt is answered, the summary is presented. The animated agents also provide feedback that can be positive, negative, or neutral ("yes", "no", "OK"). Dr. Quinn sometimes asks for more information, called "pumps" ("what else?"), or corrects misconceptions given by Glass or the human player.

The agent who delivers the different types of information is dictated by the agent's conversational role: teacher, student, and bystander. In the teaching trialog, the human player serves as the teacher, Glass as the student, and Dr. Quinn as a knowledgeable bystander. Consequently, Glass (as the student), asks the player (who presumable is the expert on this topic) the primary question. Because Glass is the one who is seeking clarification, he poses hints and prompts that are phrased as believably sounding bits of partial knowledge. Also because Glass is being taught by the player, he is the one who gives the summary (this also provides 'proof' that he now understands having been taught by the player). In the standard trialog format, the question, hints, prompts and summary are all given by Dr. Quinn directed toward the player, and some feedback is given by Glass, when he asserts that "he doesn't understand." In the vicarious trialogs, the player fulfills the bystander role who listens to Dr. Quinn tutor Glass. To maintain engagement with the material, the player is always asked whether he or she thinks Glass understands the concept based on his answers.

The conversational management, feedback and natural language understanding that the trialogs require is based on AutoTutor (Graesser, et al., 2004; Graesser, McNamara, & VanLehn, 2005; Graesser, Person, & Harter, 2001; Graesser, Wiemer-Hastings, Wiemer-Hastings, & Kreuz, 1999). AutoTutor is an Intelligent Tutoring Sytem (ITS) that helps students learn a domain by holding conversational dialogs between the student and an animated pedagogical agent. AutoTutor has brought about considerable learning gains comparable to one-on-one

human tutoring (Graesser, Chipman, Haynes, & Olney, 2005; Graesser, et al, 2001; VanLehn, et al., 2007).

Versions of AutoTutor have been built to teach computer literacy (Graesser et al., 2004) and Physics (VanLehn et al., 2007). The program simulates human tutorial dialogues in natural language. It was constructed from extensive research on human tutorial dialogs (Graesser, Magliano & Person, 1995), constructivist theories of learning (Aleven & Koedinger, 2002), and other intelligent tutors that adapt to the learner at a fine-grained level (VanLehn, et al., 2007).

Conversations in AutoTutor are largely governed by a curriculum script, which provides for each scenario (e.g., a question) an ideal answer, a set of expectations (content that the tutor would like to be expressed by the learner), a set of hints and prompts for each expectation, misconceptions and corrections, and a summary. AutoTutor will pose a scenario (e.g., a question) and the learner's answer will be assessed against the ideal answer and the expectations to indicate which of the expectations have been adequately answered or "covered". This assessment usually involves a combination of latent semantic analysis (a statistical method for representing semantic similarity between two sets of words, Landauer & Dumais, 1997) and semantic matching algorithms that consider words, word stems, and combinations of these linguistic units in regular expressions. These techniques output a numeric value indicating the semantic overlap between the student's input and the ideal answer or expectation. If the value exceeds a predetermined threshold, then AutoTutor declares that the expectation is covered by the student. If not, it will give hints to the student in order for him or her to express the content of the full expectation.

Learning Principles and the Training Module

Table 3 summarizes the links between the learning principles and features of the Training module (as well as the other two modules). The learning activities are based on the eBook and the trialogs. One important feature of the trialogs is that they are adaptive to the knowledge exhibited by the player: low, medium, and high levels of knowledge are linked with vicarious, standard, and teaching trialogs. The theoretical foundation for this linking was guided by Vygotsky's zone of proximal development in addition to the general tenet of constructionism that knowledge is actively constructed by the learner.

When prior knowledge is low, it is difficult for the learner to ask, understand, and answer questions using the desired vocabulary, so standard and teaching trialogs would be out of their "zone." Although observational learning of tutorial sessions does enhance learning (Craig, Chi, & VanLehn, 2009), there is some evidence that low prior knowledge participants show greatest learning gains when they watch a tutorial conversation, occasionally commenting on what is being learned, rather than participating directly (Craig, Sullins, Witherspoon, & Gholson, 2006). When players show a high level of knowledge, watching a tutorial dialog might be relatively boring and would result in few learning gains. Their level of knowledge enables them to be successful *active* participants. In fact, "playing teacher" is particularly effective in learning engagement and learning, as noted earlier in the context of self-explanation as a learning principle. Some learning environments contain "teachable agents" that require the human student to teach a computerized agent (Biswas et al., 2005). One called "Betty's Brain" substantially increased learning gains, transfer and self-regulated learning compared to control conditions (Biswas, Jeong, Kinnebrew, Sulcer & Roscoe, 2010). The 'teaching trialog' encompasses some of the features associated with teachable agents, namely that the users believe

that they are contributing to the knowledge of another agent, and the users receive feedback on the imparted knowledge.

There are two other reasons for engaging the player with trialogs. First, they provide an avenue to identify and fix misconceptions held by the player. People often hold misconceptions in science that might be difficult to change (Hynd & Alverman, 1989). The program stores a log of common misconceptions for each topic. If the player (or Glass) mentions any of these, the program will identify it and Dr. Quinn will correct the misconception. Second, the trialogs provide retrieval practice in that the player must identify and use particular words in answers to questions. Roediger and Karpicke (2006) have shown that retrieval practice in testing increases learning ("testing effects").

Preliminary Research on the Training Module

Because it takes several hours to progress through the entire training modules, we have only been able to do some alpha testing on the training module to see how well students learn. Twelve students attending California universities participated across 7 week, one-hour sessions. One-third of the students attended a state university, another third a community college, and the remaining third an elite private university. A pre-post test was created to measure understanding of the concepts covered by the interactive text. The same test was used before and after students went through the interactive text. The questions were a mixture of open-ended and multiple choice. The overall test scores suggested that students learned from the training module. The post scores (M = 25.04, SD = 7.06), were significantly higher than overall pretest scores (M = 15.29, SD = 8.07), $t(11) = 7.38$, $p < .01$, effect size = 1.3, and this improvement occurred for students from each of the three institutions.

Trialogs. Do the different types of trialogs affect learning? We addressed this question by having students read and answer the sets of questions to five chapters. Immediately, and after two days, they answered an open-ended comprehension test of the concepts that were addressed by the questions. Because we were interested in examining the effect of the trialogs, we did not have the participants read the chapter. Instead, they only read and answered the multiple choice questions, and either received a vicarious trialog, a teaching trialog, or a mixture of the three based on their performance (adaptive). Hence, we were interested whether animated agents delivering formative feedback would lead to differential "testing effects" (Roediger & Karpicke, 2006).

We found that the trialogs had little impact on immediate testing but did have an impact after a two day delay. The means (the percentage correct) for the vicarious, adaptive and teaching conditions on the immediate test were .45 (*SD* = .19), .48 (*SD* = .19), and .47 (*SD* = .18), respectively. The corresponding means in the delay condition were .34 (*SD* = .16), .44 (*SD* = .19) and .45 (*SD* = .21), respectively. The drop in scores due to the delay was significant in the vicarious condition [t (27) = 4.23, $p < .01$] but not in the other two trialog conditions (p's < .40). In addition, the adaptive and teaching scores in the delayed condition were significantly greater than the vicarious condition ($p < .05$). The pattern of means indicate greater learning in the adaptive and teaching conditions compared to the vicarious conditions. We had hypothesized that the adaptive condition would outperform the teaching condition, but it did not. Rather, there was no difference between the adaptive and teaching conditions. We kept the adaptive trialogs in *Operation ARIES!* instead of only using teaching trialogs for two reasons. First, they contain a variety of interchanges that we hope players will value. Second, participants in this study only

responded to 5 chapters worth of multiple choice items without having read the eBook. Perhaps other patterns of results would be found when participants read the entire eBook.

Module 2: Case Studies

In this module, the player reads and evaluates a number of brief research reports. Each report describes a study in one of the following domains: psychology, biology, or chemistry. Each is written and formatted in such a way that it resembles newspaper or magazine article, a blog, a web page, or an advertisement. What they have in common is that virtually all contain one or more flaws pertaining to the concepts taught in the Training module. For example, a research report might not include a control group, not have a valid dependent variable, or might suffer from experimental bias. The purpose of the module is to teach the player how to evaluate research reports (called case studies) by having them identify flaws contained in them.

We adopted a case-based learning environment for flaw identification because cases (problems, instances, scenerios) allow learners to encode and discover the rich source of constraints and interdependencies underlying the target elements (flaws) within the cases. The memory of prior cases provides a knowledge base for assessing new cases, in that they help guide reasoning, problem solving, interpretation and other cognitive processes. That is, players implicitly learn the various ways in which the flaws are instantiated in different contexts, and how they are causally and conceptually connected. For example, if a study suffers from biased sample selection, then it would likely suffer from poor generalizability since its finding was based on a sample with particular characteristics that may not occur in the larger population.

Table 4 presents a sample “case” along with the flaws that it contains, whereas Figure 2 presents a screen shot of the case studies interface. There is a picture representing the human

player (e.g., "Gary"), as well as two new agents joining Dr. Quinn: Tracy and Broth. Tracy, like the player, is a fellow student agent at the FBS. The human player and Tracy compete against each other for the honor of going forward to the next module where they will interrogate aliens. Broth is an alien defector who wants peace with Human "Beans" and who is observing the sessions. (Glass Tealman, who was the fellow student during the Training module, left at the end of that module to find his older brother who is being held captive by the Fuaths.) As in the training module, Dr. Quinn provides instructional support by giving guidance and feedback to each player (the human player and Tracy). Broth gives feedback to the players as well, but also advances the story line by providing knowledge regarding the Fuath's perspective on their spying mission. During the course of the case evaluations, Broth announces that the Fuath spies communicate with each other by placing specific flaws in the research that they publish. By deciphering the flaws, Broth reports several developments pertinent to the story line, the most distressing being that the very deadly aliens called Nemotoads are en route to Earth. This news is met with some disbelief within the FBS, creating uncertainty and tension.

The human player and Tracy take turns evaluating cases. The current player (human or Tracy) first chooses a case to evaluate from a list, reads it, and then types in flaws into an input box. For support, the player can buy a list of flaws or read the "Big Book of Science." After the flaw is entered, the program matches the input to the list of flaws using key word matching algorithms. The algorithm involves computing a match score between the input and each flaw. If the score falls below a threshold, Dr. Quinn will ask the player to rephrase the flaw; otherwise the flaw with the highest match score will be presented in the "closest match" box. When the player is satisfied with the match, the player requests the answer. Dr. Quinn then gives elaborative feedback as to why the answer was correct or incorrect, and points are either added to

or subtracted from the player's current score. If correct, the player retains his or her turn, and is asked to identify another flaw. The player can also push the "No (more) Flaws" button if he or she believes all flaws have been identified (or if there were no flaws in the case in the first place.) If the player is incorrect, then the turn passes to the other player. Periodically, Dr. Quinn asks the player to justify his or her answer by asking "Why did you choose that flaw?" in order to encourage self-explanation and to discourage random answers.

If there are any unidentified flaws left by the time that both players press the "No (more) Flaws" button, then Dr. Quinn provides a brief tutoring session with the current player. The dialog uses a curriculum script similar to the ones used for the "standard" trialogs described earlier. The program selects a hint associated with a flaw that has not been covered, and it is delivered by Dr. Quinn. If the hint is unsuccessful in eliciting the desired response (e.g., there is no control/comparison group), then Dr. Quinn gives an appropriate prompt. Table 4 shows some of the hints and prompts. If the player fails to answer this question, then the opponent has the opportunity to answer for maximum points. In regard to points, points are awarded on the basis of the presence and type of question: without hint → hint → prompt.

Learning Principles and the Case Studies Module

One prominent feature of the case studies is that the cases describe a variety of content (psychology, biology, and chemistry) written and formatted in a number of ways. They read and look like blogs, advertisements, and newspaper and magazine articles. The variation in content and format was designed to promote transfer - to use scientific inquiry skills in other contexts. In addition, there are many types of feedback given to the player. When a player types in a flaw, he

or she receives immediate feedback as to whether that flaw is present in the case. The feedback is presented verbally by Dr. Quinn and by the addition or subtraction of points.

One important feature not listed in Table 3 is competition. The player is competing against Tracy. Competition is often listed as a characteristic of games, both digital and nondigital (Yee, 2006). A survey conducted by the Annenberg School for Communication Games Group at the University of Southern California indicated that competition was the primary reason why players chose certain games (as cited in Bryant & Fondren, 2009). However, some designers caution against competition in favor of cooperation because it may focus on the act of winning rather than on the learning domain (Whitton, 2010). One limitation of the use of competition in Operation ARIES! is that it may not benefit all groups in the same way. For example, boys tend to choose more competitive games than girls (Hartman, 2003), so competing against Tracy might not be attractive to all players.

Research on the Case Studies

The research that we have conducted using a precursor to the case studies module suggests that having students engage in tutorial dialogs about case studies is an effective strategy in learning how to identify flaws. In a study conducted by Kopp, Britt, Millis and Graesser (in press), undergraduate psychology students listened to an animated teacher agent read several research descriptions used in the module, and immediately after each, the participates were assigned to one of three activities: (1) listen to a conversation between the teacher and an animated student in which the participant had to write down the flaws that were identified and summarized by the teacher, (2) write down flaws that the participant noticed before the teacher provided a summary of the flaws, or (3) participate in a tutorial dialog with the teacher agent.

The tutorial dialogs were similar to the standard trialogs but without a third agent. Learning gains were assessed by comparing pre- and post-test scores on task that required the participants to critique other flawed studies. When the post-test scores were adjusted for the pre-test scores, Kopp et al. (in press) reported significantly greater learning when participants had participated in full dialogs than when they listened to and wrote down the flaws (Experiment 2). Interestingly, they found that engaging in tutorial dialogs was not necessary for all cases to maintain the advantage. The highest rates of learning occurred when one-half of the cases required full dialogs (condition 3 above) and the other half required an initial answer (condition 2 above). This outcome is somewhat counterintuitive, but makes sense after some reflection. Although dialogs increase learning, it appears that it is most efficient to only have the participants engage in dialogs for one-half of the cases. The extra dialogs might incur fatigue. We used this finding to inform the design of Case Studies in which the human player directly evaluates only one-half of the cases whereas Tracy evaluates the other half.

In another study, we compared learning and reactions to the Case Studies module (game condition) with a version that lacked points, competition, and animated agents (non-game condition). Otherwise, the same materials and feedback were administered in the two conditions. As in Kopp et al. (in press), participants were undergraduate psychology students, and were given a pre- and post-test that required them to correctly identify flaws in research cases. There were four one-hour sessions that occurred across two weeks. Immediately after the first and last sessions, we asked participants about their level of engagement, motivation, interest, challenge, and frustration using a 6-point Likert-type scale. As a measure of the ability to detect flaws, we computed a "flaw identification score" by subtracting the participant's "false alarm" rate (i.e., the percentage of occurrences when the participant said a flaw was present in the research but it was

not) from their "hit" rate (i.e., the percentage of occurrences when a participant correctly identified the presence of a flaw). A score of zero on the "flaw identification score" would indicate no ability for a person to discriminate between the presence and absence of a flaw, whereas a score of 1.0 would indicate perfect discrimination. The pre-test means for the game (M = .09) and no game (M = .07) conditions were low and nearly identical. The increase on the post-test scores was significantly higher in the game (M = .43) than in the no game (M = .32) condition, as indicated by a significant interaction, $F(1, 27) = 3.01$, $p < .05$, $MSe = .012$ (one-tailed). Interestingly, few differences emerged on the questions about their experiences. Participants in the game conditions gave significantly higher ratings on interest ($p < .05$, one-tailed) and frustration ($p < .01$), but there was no significant difference on the other measures. The finding for frustration was unexpected, but holding an informal focus group afterwards was revealing. The participants in the game condition expressed some frustration from having to use the desired vocabulary (e.g., no control group, small sample size) required for a successful match between the user input and the flaw categories. However, they expressed even more frustration by the time taken up by Tracy's responses which appeared as text being typed in real time. Since then, Tracy's text responses on the screen have been dramatically sped up.

Module 3: Interrogation

The story line heats up in this last third of the game. Thousands of Nemotoad spacecrafts have left the Mother ship and have settled into geocentric Earth orbits. An intercepted message reveals their horrific plans: the Grand Nemotoad will order humans to be subjugated as slaves and they will scorch the Earth. In an attempt to capture the Grand Nemotoad and other aliens, the FBS has conducted a large-scale raid, arresting dozens of scientists suspected of being aliens. However, some of those who were arrested are human. Only through interrogating each suspect

on their research can the FBS know for sure the species of each suspect. The aliens are those publishing flawed research, the humans are not. Meanwhile, the clock is ticking toward global disaster.

The Interrogation module teaches the player how to evaluate research by asking questions. The player first reads a summary of research conducted by one of the suspected alien spies. Similar to case studies, the research is presented in different types media (newspapers, blogs, etc.), but unlike the case studies the research is abbreviated and critical information is missing. The description of the research might be the length of an abstract (roughly 150 words), an advertisement, and in some cases, it might be only a headline (e.g., "Study shows music helps plants to grow"). The descriptions do not explicitly signal any flaw. Hence, in order to uncover a flaw, the player must ask the suspect questions about the research, and classify each answer on whether it revealed a flaw or not. If the study is flawed, then the suspect should be judged an alien. If there is no major flaw, or if the suspect acknowledges a flaw found in the study, the suspect should be judged human.

Figure 3 shows a screen shot of the Interrogation module after the research description has been read. In addition to Dr. Quinn, there are two other agents: Scott, serving as the interrogator, and the suspect who is hidden behind a screen. Scott and the suspect are located in the same room, apart from the player, and the player is responsible for sending Scott questions to ask the suspect. The player sends a question by typing it into an input box. Having the player send questions to the suspect via Scott solved a technical problem that could arise if the player and suspect communicated directly. In that circumstance, the program might misclassify a question typed in by the player. If this were to happen, the suspect would answer a question not posed by the player, leading the player to be confused. However, we had the player send

questions to Scott, and Scott hedges when there is a low match score between the question typed in by the player and a stored question (e.g., "I think I know what you are getting at." "Your input is coming in fuzzy – I think I heard you"). The hedge allows Scott to ask any question (regardless of the input match) and receive an answer that does not sacrifice the buy-in that Scott truly understands the player's questions.

After the player receives an answer, the player evaluates the answer by checking off options (flaw, flaw recognized, no flaw) on relevant subcategories on a "score card." There is a total of 25 subcategories, arranged under the following superordinate categories: hypothesis, independent variable, dependent variable, control, sample, experimenter, conclusions. For example, the subcategories of "control" are possible confounds, subject bias, control groups, random assignment, mortality and attrition. The player might be able to correctly identify the answers to more than one subcategory from a single answer. For example, consider the following question and answer about the dependent variable (outcome variable) in the context of study that was conducted to test whether an advertised video decreases shyness:

Question: "Tell me about how you defined shyness."

Answer: "I used a shyness scale that I came up with. What you do is count the number of times a person blinks as they talk to you. It is quite reliable. People tend to be consistent in the number of times they blink per minute."

A correct classification of this answer would be "flaw" on the subcategory "validity" because eye blinks are not valid indicators of shyness, and "no flaw" on "reliability" because counting eye blinks is likely to be reliable. Both "valid" and "reliable" are subcategories under dependent variable.

The answer category "flaw recognized" is relevant when the suspect recognizes that there is a flaw or limitation. For example, consider the following question and answer about how the researcher selected participants (sample selection):

Question: "How did you get your participants?"

Answer: "The participants were attendees for a seminar designed to reduce shyness. We had advertised on radio and local television. Oh, wait, this is called selection bias, right? I guess this was a problem."

A sample selection flaw occurs in psychological research when the participants select themselves for the study. When this occurs, the participants might be biased in any number of ways, decreasing the likelihood that the results would generalize to other participants. Therefore, a correct classification would be "flaw recognized" under the subcategory "sample selection" because the researcher acknowledged the problem. The player is also instructed that because of their training, aliens never lie.

The "scorecard" contains a number of design elements that provide for instructional support. These include 'clickable' subcategories, which when clicked, Scott asks a question based on that subcategory. The scorecard highlights categories and subcategories that could be addressed by the answer. These features are systematically taken away as the player advances through the module, resulting in four difficulty levels. At the end of the module, the scorecard itself is taken away, and the player is asked to make the "alien/human" judgment as soon as a flaw is revealed. Points are awarded for correct responses and taken away for incorrect responses. The point values vary based on the difficultly level of the game. At the end of the question-answer pairs for a given case, the player is asked to judge whether the suspect is human

or alien. If there is at least one flaw, the correct judgment is alien. Otherwise, the correct judgment is human.

Learning Principles and the Interrogation Module

The primary learning principle that this module incorporates is question-asking, a learning principle not covered in Table 3. Prior research that has indicated learning is increased by formulating questions, and that students can learn how to ask good questions (Beck, McKeown, Hamilton, & Kucan, 1997; Graesser, Ozuru & Sullins, 2009; King, 1994; Palincsar & Brown, 1984; Taboada & Guthrie, 2006). Asking questions increases learning not only by receiving answers to the posed questions, but in order to ask a question, the learner must actively read the material, encoding it deeply, and regulate their own learning in the process (Wisher & Graesser, 2007). Rosenshine, Meister and Chapman (1996) reviewed the literature on teaching question-asking strategies while reading and report mean effect sizes of .36 and .86 (in standard deviation units) when standardized and experimenter-generated tests of comprehension were used. Increasingly, learning environments are incorporating question facilities that encourage users to ask open ended questions (Linn, Davis & Bell, 2004; Palincsar & Brown, 1984).

Concluding Comments

Serious games lay at the intersection among content (e.g., research methods), game design (e.g., dialogs, agents, story), and pedagogical theory (e.g., principles of learning and motivation). Ideally, a serious game should be fun to play and educational. This sweet spot is notoriously difficult to achieve. One reason for this difficulty is that a game's enjoyableness and fun depend on a number of factors whose interactions are not clearly understood. Some commonly cited dimensions of enjoyable gaming experiences are overall game design, aesthetic

(visual and auditory) presentation, the ease and effectiveness of control, complexity/challenge, social interaction/community, and storyline/narrative (Wang, Sehn & Ritterfeld, 2009; Whitton, 2010). Ensuring that a game contains these characteristics is monetarily expensive, but also varies based on a player's individual tastes. Second, enjoyable game play experiences may not translate into deep learning (Graesser, Chipman, Leeming, & Biedenbach, 2009). Deep learning of a complicated content or skills will probably require deep cognitive and emotional investments that span many hours of practice which not be considered fun by the player. Like many educational and serious games, *Operation ARIES!* tries to initiate and maintain interest by immersing the player into a larger narrative (Ratan & Ritterfeld, 2009). We have had many positive responses to the storyline of *Operation ARIES!* when presented alone, but it is currently unknown whether it can overcome negative reactions (e.g., boredom, frustration) that might arise from practice and repetition required to learn critical thinking skills in science. One high school student gamer who is skeptical regarding the value of educational games put it this way: "Educational games are homework, and I can smell homework!"

Operation ARIES! is in its final stages of empirical testing and these concerns are currently being explored. We are hopeful that the program will be valued by both students and educators. Pearson Education plans to host the game on a website where students can login and play, and where instructors can monitor their students' progress. Students can play at their leisure, and instructors can incorporate the game into their course in a number of ways. Some instructors might assign it as homework and only care whether the students complete the game, whereas others might incorporate the assessments which are internal to the game into the student's course grade. Still other instructors might only assign some modules and not others. It is anticipated that this level of flexibility will make *Operation ARIES!* a desirable addition to

scientific learning in traditional classrooms as well as in distance learning. We hope that you have enjoyed getting to know the science behind *Operation ARIES!* as much as we have enjoyed sharing it with you.

Authors Note

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B070349 to Northern Illinois University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Table 1

Scientific inquiry topics covered by Operation ARIES!

Theories and the Experimental Method
Hypothesis
Science and Pseudoscience
Operational Definitions
Independent Variables and Participant Variables
The Dependent Variable
Dependent Variables: Reliability, Accuracy, and Precision
Dependent Variables: Validity
The Dependent Variable: Objective Scoring
Replication of Results
Experimental Control
Control Groups
Random Assignment to Groups
Subject Bias
Attrition and Mortality
Representative Samples
Sample Size
Experimenter Bias
Conflict of Interest
Causality vs. Correlation

Drawing Conclusions: Generalizeability

Table 2

Examples of teaching, standard, and vicarious trialog tutorial sessions.

Teaching Trialog (importance/function question for "operational definition")

Glass:	Okay, I know operational definitions are important in research, but I'm not sure why. Let's just go over the importance of these things one more time. Why do we need to have operational definitions? (question)
Laura:	so that they can be reliabily measured and used by other researchers
Glass:	OK. (neutral feedback)
Dr. Quinn:	Um, can you add to that? (pump)
Laura:	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. I thought operational definitions had something to do with telling others about your study. What could that be? (Hint)
Laura:	how to measure a variable.
Glass:	I still don't understand. Wow, it seems like operational definitions have something to do with being able to reliably do something . What is that something? (prompt)
Laura:	measure a variable
Dr. Quinn:	Okay. Reliably recognize variables (neutral feedback, prompt completion)
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? (prompt)
Laura:	reliable
Dr. Quinn:	Great! It is reliable (positive feedback, prompt completion)
Glass:	Yeah, I understand. Operational definitions are important to consider because they allow a particular variable to be reliably recognized, measured, and understood by all researchers. (summary)

Standard Trialog ('example' question for "subject variable")

Dr. Quinn: I just want to make sure that everyone really knows this. Why don't you explain why the answer about socioeconomic status is an example of a subject variable?

Lauren: so that they are getting a generalization for their experiment

Glass: I do not understand.

Dr. Quinn: Think about this. As you're considering why the socioeconomic status option is the correct answer, think about how it mentions classifying the participants. Now try to explain why socioeconomic status provides a good example of a subject variable. (hint)

Lauren: so that they can have the same socioeconomic status

Glass: still don't understand.

Dr. Quinn: Okay. Let's try this. In this example, the researcher is using a characteristic about the subjects to put them into what? (prompt)

Lauren: a specific category

Dr. Quinn: Okay, into groups. I bet you know the answer to this. In this example, the researcher is using socioeconomic status to group the subjects, and for the subjects, socioeconomic status is a what? (prompt)

Lauren: specific group

Dr. Quinn: Alright. Subject variables are characteristics of the subjects, such as socioeconomic status, that are used to categorize them into groups. They are a type of independent variable that is not manipulated. (summary)

Vicarious Trialog (importance/function question for "theory")

Glass: You know, I thought the correct answer was the methodologies. I guess I need some help here.

Dr. Quinn: Explain why you think theories are important.

Glass: A theory provides an understanding.

Dr. Quinn: Okay. Here is a little hint. As you consider why theories are important, think about how theories might relate to how research is conducted and what the research findings are. (hint)

Glass: They provide predictions for future research projects.

Dr. Quinn: Tyler (human player), what is your opinion? Does it seem like Glass now understands why theories are important? Please answer "Yes" or "No".

Tyler: No.

Dr. Quinn: You are wrong, Tyler. Theories are important in research because by providing explanations for findings, they are able to organize many research findings and make predictions for future research projects. (corrective feedback, summary)

Table 3

Learning Principles and the Three Modules.

	Modules		
Learning principle	Training	Case Studies	Interrogation
Zone of proximal development	Type of trialog is based on prior knowledge.	N/A	Scaffolding of "score card", progressively more difficult
Self-explanation	Reflection questions are posed in eBook; trialogs require player to explain concepts and answers.	Quinn periodically asks player to justify the selection of a flaw.	N/A
Feedback	Agents give corrective and elaborative feedback in the trialogs.	Points are awarded for correct answers; points are correlated with less scaffolding (no hint → hint → prompt); summary of case is presented.	Points, corrective feedback, summary of case is presented.
Narrative, Fantasy, Adventure	The eBook is an alien spy book and the player is training to be a special agent of the FBS.	E-mails, dialogs, and deciphered flaws advance story line. Nemotoads threaten world.	Story is advanced through news updates, and emails. World is saved.
Player Control	Player can opt out of reading by taking "challenge test"	Player chooses case to evaluate, can buy "flaw list" or access the eBook.	Player selects cases and difficulty level.
Dialogue	Player engages in tutorial dialogs called trialogs.	Dr. Quinn gives hints and prompts, similar to the "standard" trialogs.	N/A
Encoding variability	Novel examples and illustrations are presented in the eBook	Research is written in a number of formats; topics in psychology, biology and chemistry.	Research is written in a number of formats; topics in psychology, biology and chemistry.

Table 4

Sample Case Study, Hints and Prompts

Title: The Battle of the Sexes

Who are more aggressive, men or women? Popular media and news reports often portray men as the more aggressive gender. But think again: there was *Bonnie* in the infamous Bonnie and Clyde, and of course, the movie *Mean Girls*. Perhaps the genders are closer on aggression than one might think.

Dr. Alan Maye, a member of a research institution which focuses on aggression, wanted to find out if one gender is really more aggressive than the other.

To conduct his experiment, he placed an ad in a newspaper and asked for volunteers to participate in a study that was going to explore gender differences in aggression. All interested people were asked to report to the institution conducting the research.

Twenty-five men and twenty-five women volunteered to participate in the study. When they arrived at the study, they were exposed to multiple situations that were supposed to elicit aggression (an accomplice posing as a participant was used to provoke the actual participants). After they were put in this situation, the participants were given the opportunity to write a message to the person who provoked them.

The messages that participants wrote were coded by two independent researchers who were not aware of the participant's gender. The messages were coded on a 7-point scale for the degree of verbal aggressiveness that was used. The results showed that women provided more aggressive messages than men.

In a follow-up study, the researchers found the same results—women were found to be more aggressive than men. Based on these results, the researchers concluded, contrary to popular belief, women are actually more aggressive than men.

Flaw: dependent variable is not valid

Hint: Considering that verbal aggressiveness is simply one type of aggressive behavior, what can you say about the dependent measure used here?

Prompt: A dependent variable that measures something other than what it is claimed to measure is called what?

Flaw: poor sample selection

Hint: What flaw is associated with how Dr. Maye chose participants for this study?

Prompt: Because this study only included participants who were interested in answering the advertisement, the study involved a poor selection of the what?

Flaw: subject bias

Hint: What can you say about the fact that the newspaper ad informed possible participants of the intent of the study?

Prompt: If participants can influence results based on their own expectations regarding the experiment, this is what type of bias?

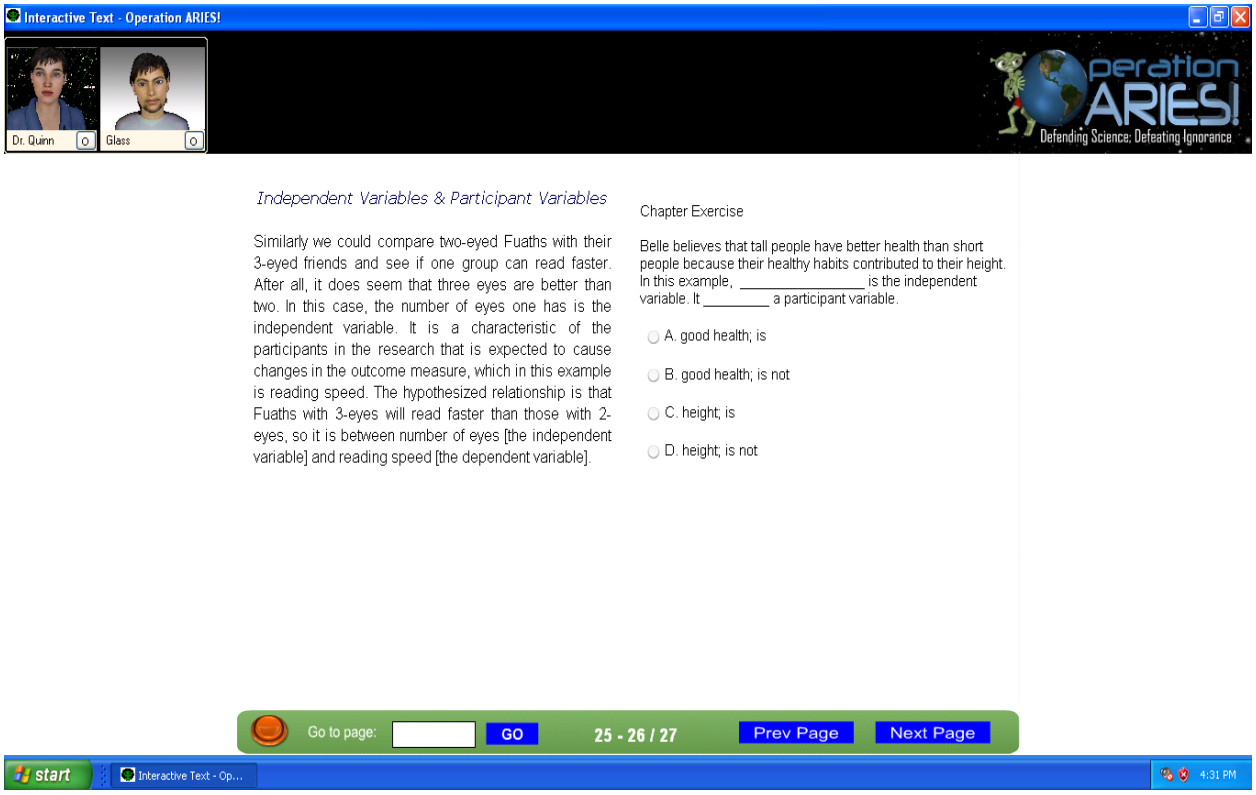

Figure 1. Training Module.

Figure 2. Case Studies

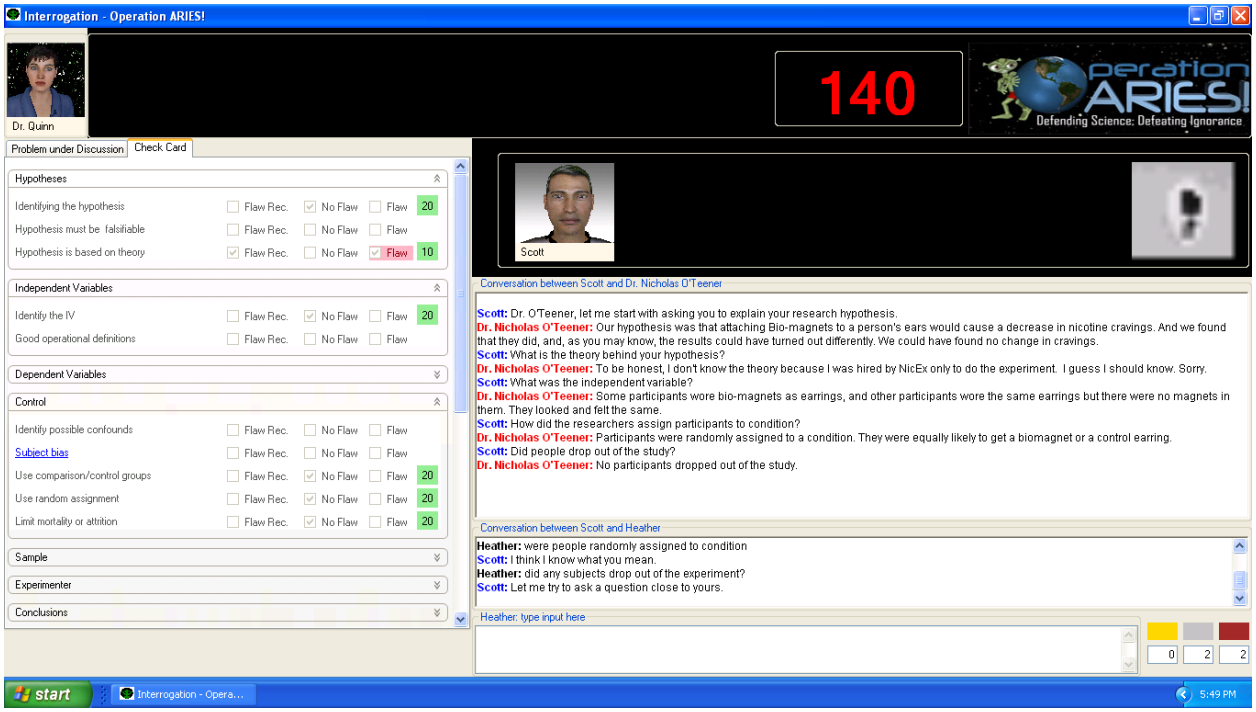

Figure 2. Interrogation module.

References

Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science, 26*, 147-179.

Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*, 167-207.

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology, 94*, 416-427.

Baylor, A. L. & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education, 15*, 95-115.

Beck, I.L., McKeown, M.G., Hamilton, R.L., & Kucan, L. (1997). *Questioning the author: An approach for enhancing student engagement with text*. Delaware: International Reading Association.

Benyon, D., Turner, P., & Turner, S. (2005). *Designing Interactive Systems*. Harlow: Addison-Wesley.

Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & The Teachable Agents Group at Vanderbilt. (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artifical Intelligence, 19*, 363-392.

Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., & Roscoe, R. (**2010).** Measuring Self-regulated Learning Skills through Social Interactions in a Teachable Agent Environment**.** *Research and Practice in Technology-Enhanced Learning, 5*(2), 123-152.

Bloom, B. S. (1956) *Taxonomy of Educational Objectives, the classification of educational goals – Handbook I: Cognitive Domain*. New York: McKay.

Bransford, J.D., Sherwood, R.S., Hasselbring, T.S., Kinzer, C.K., & Williams, S.M. (1990). Anchored Instruction: Why we need it and how technology can help. In D. Nix & R. Spiro (Eds.), *Cognition, education, and multimedia: Exploring ideas in high technology* (pp. 115-141). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bryant, J. & Fondren, W. (2009). Psychological and communicological theories of learning and emotion underlying serious games. In U. Ritterfeld, M. Cody, and P. Vorderer (Eds.) *Serious games: Mechanisms and effects* (pp. 103-116).

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471-533

Craig, S. D., Chi, M. T. H., VanLehn, K. (2009). Improving classroom learning by collaboratively observing human tutoring videos while problem solving. *Journal of Educational Psychology*, 101, 779–789

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). Deep level reasoning questions effect: The role of dialog and deep-level reasoning questions during vicarious learning. C*ognition and Instruction, 24*, 563–589.

Csikszenthimhalyi, M. (2002). *Flow: The Psychology of Happiness*. Londun: Ranom House.

Gee, J.P. (2003). *What video games teach us about language and literacy*. New York: Palgrave/Macmillan.

Graesser, A.C., Chipman, P., Haynes, B.C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education.*

Graesser, A.C., Chipman, P., Leeming, F., & Biedenbach, S. (2009). Deep learning and emotion in serious games. In U. Ritterfeld, M. Cody, and P. Vorderer (Eds.) *Serious games: Mechanisms and effects* (pp. 83-102).

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

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

Graesser, A.C., McNamara, D.S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist, 40*, 225-234.

Graesser, A., Ozuru, Y., & Sullins, J. (2009). What is a good question? In M. G. McKeown & L. Kucan (Eds.), *Threads of coherence in research on the development of reading ability* (pp. 112-141). NY: Guilford.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, *9*, 359.1-28.

Graesser, A. C., Person, N., Harter, D., & the Tutoring Research Group (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education, 12,* 257-279.

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the TRG (1999). Auto Tutor: A simulation of a human tutor. *Journal of Cognitive Systems Research, 1,* 35-51.

Griffin, T. D., Wiley, J. & Thiede, K. W. (2008) Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36,* 93-103.

Halpern, D. F. (2002). Teaching for critical thinking: A four-part model to enhance thinking skills. In S. Davis & W. Buskist (Eds.), *The teaching of psychology: Essays in Honor of Wilbert J. McKeachie and Charles L. Brewer* (pp. 91-105). Mahway, NJ: Lawrence Erlbaum.

Harackiewicz, J. (1979). The effects of reward contingency and performance feedback on intrinsic motivation. *Journal of Personality and Social Psychology, 37,* 1352–1363.

Hartmann, T. (2003). Gender differences in the use of computer-games as competitive leisure activities. Paper presented at *Digital Games Research Association* (DIGRA), November 4-6, 2003, Utrecht, The Netherlands.

Heeter, C., Egidio, R., Punya, M., Winn, B., & Caywood, J. (2007). Alien games: Do girls prefer games designed by girls? *Games and Culture*, Vol. 4, No. 1, 74-100 (2009) DOI: 10.1177/1555412008325481.

Hynd, C. & Alverman, D.E. (1989). Overcoming misconceptions in science: An on-line study of prior knowledge activation. *Reading Research and Instruction, 84*, 12-26.

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31*, 338-368.

Kopp, K., Britt, A., Millis, K., & Graesser, A. (in press). Improving the efficiency of dialogue in tutoring. *Journal Learning and Instruction*.

Kulik, J. A., & Kulik, C-L. C. (1988). Timing of feedback and verbal learning. Review of Educational Research, 58, 79-97.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Linn, M. C., Davis, E. A., & Bell, P. (2004*). Internet environments for science education*. Erlbaum: Hillsdale, New Jersey.

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

Malone, T.W., & Lepper, M.R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In R.E. Snow & M.J. Farr (Eds.), Aptitude Learning and instruction, 3 (pp. 223-253). Hillsdale, NJ: Erlbaum.

McNamara, D., S., O'Reilly, T., Rowe, M., Boonthum, C., & Levinstein, I., (2007). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies (pp. 397-420). In D.S. McNamara (Ed.*) Reading comprehension strategies: Theories, interventions, and technologies*. New York: Erlbaum.

Meyer, B.J.F. & Wijekumar, K. (2007). A web-based tutoring system for the structure strategy: theoretical background, design, and findings (pp. 347-374). In D.S. McNamara (Ed.*) Reading comprehension strategies: Theories, interventions, and technologies*. New York: Erlbaum.

National Science Education Standards (NSES). (1996). Washington, DC: The National Academies Press:

Oxland, K. (2004). *Gameplay and design*. Harlow: Addison-Wesley.

Palincsar, A. S., Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension- monitoring Activities. *Cognitive and Instruction, 1*,117-175.

Ratan, R. & Ritterfeld, U. (2009). Classifying serious games. In U. Ritterfeld, M. Cody, and P. Vorderer (Eds.) *Serious games: Mechanisms and effects* (pp. 10-24).

Rieber, L. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Education and Technology Research & Development, 44*, 42-58.

Roediger, H. L., & Karpicke, J. D. (2006) Test-Enhanced Learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249-255.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research, 66,* 181-221.

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game Design and Fundamentals.* The MIT Press. Cambridge, Massachusetts.

Shute, V. J. (2006). Focus on formative feedback (ETS Research Report). Princeton, NJ: ETS.

Taboada, A. & Guthrie, J.T. (2006). Contributions of student questioning and prior knowledge to construction of knowledge from reading information text. *Journal of Literacy Research, 38,* 1-35.

Van Eck, R. (2007). Building Intelligent Learning Games. In D. Gibson, C. Aldrich, & M. Prensky (Eds) *Games and Simulations* in *Online Learning Research* & *Development Frameworks.* Hershey, PA: Idea Group.

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

Vygotsky, L. (1978). Mind in Society: *The Develop,ment of Higher Psychological Functions.* Cambridge, MA: Harvard University Press.

Wang, H., Shen, C., & Ritterfeld, U. (2009). Enjoyment of digital games: What makes them "seriously" fun? In U. Ritterfeld, M. Cody, and P. Vorderer (Eds.) *Serious games: Mechanisms and effects* (pp. 25-47).

Whitton, N. (2010). *Learning with digital games: A practical guide to engaging students in higher education*. New York: Routledge.

Wisher, R. A. & Graesser, A. C. (2007). Question asking in advanced distributed learning environments. In S.M. Fiore and E. Salas (Eds.), *Toward a science of distributed learning and training*. Washington, D.C.: American Psychological Association.

Yee N. (2006). Motivations of play in online games. *CyberPsychology & Behavior, 9,* 772–775.

Yeomans, J. (2008). Dynamic assessment practice: some suggestions for ensuring follow up. *Educational Psychology in Practice, 24(2),* 105-114.