Tech Know Learn (2017) 22:1–22
DOI 10.1007/s10758-016-9289-5

ORIGINAL RESEARCH

The Impact of Game-Like Features on Learning from an Intelligent Tutoring System

Keith Millis[1] · **Carol Forsyth**[2] · **Patricia Wallace**[1] · **Arthur C. Graesser**[3] · **Gary Timmins**[1]

Published online: 24 June 2016

Abstract Prior research has shown that students learn from Intelligent Tutoring Systems (ITS). However, students' attention may drift or become disengaged with the task over extended amounts of instruction. To remedy this problem, researchers have examined the impact of game-like features (e.g., a narrative) in digital learning environments on motivation and learning. Some of this research has concluded that the game-like features decrease learning because the features take away resources from the primary task of learning subject-matter content. However, these experiments have involved short-term interventions of less than an hour. Two experiments using college students examined the impact of adding game-like features to the ITS AutoTutor in an intervention that lasted 4 h. In one study, a game-like version was compared to a text-only version and a "do nothing" control. In another study, a game-like version was compared to a nongame version that had similar interfaces. Unlike prior research that has shown that narratives decrease learning in digitally-based learning environments, the game-like features, which included a narrative, had little impact on learning from the ITS. Reasons for the discrepancies are discussed.

Keywords Intelligent Tutoring Systems · Serious games · Learning

1 Introduction

It is easy to imagine a classroom of students anxiously waiting for class to end so that they can race home and log into a virtual reality where all of the stress of the day seemingly disappears as they are immersed in a fantasy world. The students do not worry about the

✉ Keith Millis
kmillis@niu.edu

[1] Northern Illinois University, Dekalb, IL, USA

[2] Educational Testing Service, Princeton, NJ, USA

[3] University of Memphis, Memphis, TN, USA

upcoming science, chemistry or physics tests because instant gratification is at their fingertips with games such as *World of Warcraft* (Blizzard Entertainment 2004) or *The Covenant*. Meanwhile, the scientific literacy rates seem to stagnate in mediocrity. For example, in 2009, the National Assessment for Educational Progress reported that only 21 % of 12th graders in the United States reached a proficient level in science (Institute of Education Sciences 2009). Furthermore, in 2010, the National Science Foundation reported that only 42 % of Americans show an understanding of the scientific inquiry process (National Science Foundation 2012).

Cognitive psychologists, computer scientists, and many other experts within the walls of the ivory towers and research facilities try to keep up with the need for better education while still trying to make learning "fun." The phenomenon can be seen with the emergence of a multitude of computerized serious games that teach across the educational spectrum (Richards et al. 2013). Topics have included microbiology (Rowe et al. 2011), electromechanical devices (Koening 2008), research methods (Halpern et al. 2012; Millis et al. 2014), and reading comprehension (Jackson and McNamara 2013), just to name a few. This trend has come in part from an initial strike of scientists and developers who create Intelligent Tutoring Systems (ITS) that provide students with one-on-one education that adapts to each student's pedagogical strengths and weaknesses.

1.1 Research Questions

With these developments, two research questions become relevant. First, can we build off of pre-existing work in Intelligent Tutoring Systems when creating serious games? Second, will having these systems presented as games help increase learning, or will they have no or even negative effects? Although the primary focus is on learning, we measured some noncognitive factors as well (e.g., motivation, interest) and will address these factors while interpreting our findings on research questions 1 and 2 since these are important factors that contribute to learning (ChanMin and Pekrun 2013; Pekrun 2006; D'Mello and Graesser 2012)

1.2 Intelligent Tutoring Systems and Serious Games

There has been an increase in the development and use of ITSs in several educational settings (for review, see Graesser et al. 2012a; Woolf 2009). Many such systems include pedagogical agents to converse with students (Graesser et al. 2014). Examples of such systems teach a variety of domains, such as biology in *Betty's Brain* (Biswas et al. 2010) or *Guru* (Olney et al. 2012), mathematics in ALEKS (Nye et al. 2014), computer programming in *Coach Mike* (Lane et al. 2011), and electronics in *BEETLE II* (Dzikovska et al. 2014) or *DeepTutor* (Rus et al. 2013), to name a few.

Many of these ITSs attempt to emulate human tutors by presenting individual problems to students and by recognizing how the student attempts to solve the problems. The system notes whether or not the student is successful in answering the question and whether the student has applied the desired strategies as determined by the curriculum taught by the intelligent tutor. Underlying models of the student can diagnose strengths and weaknesses of the student based on the body of embedded questions and corresponding student answers. Based on the model of the student, the systems can choose problems intelligently, providing the student with problems that address specific short-comings of the student, along with appropriate scaffolding strategies such as hints, prompts, and different types of feedback. This is particularly appealing because the systems can provide problems or

scaffolding within the individual student's zone of proximal development (Vygotsky 1978), or the area of the optimal challenge where the material is not too easy or too difficult for the student. Finding this optimal challenge level is important for learning and motivation. These types of tactics within ITSs have been successful, as ITSs have been found to increase performance up to a similar level of human tutors (Graesser et al. 2012a; for a review see VanLehn 2011).

Although ITSs are effective in promoting learning, researchers have noted instances when participants' motivation and engagement waxes and wanes as they interact with the program over time. Indeed, like any learning environment, students may get tired, frustrated or bored from interacting with the program (Baker et al. 2010) just as they might with any type of homework assignments. Students may feel frustrated with the program if they perceive the program is not responding appropriately to their input. They may not feel that the program "understands" their responses or points of reasoning. In addition, because they know the tutor is a computer program, they may try to "game the system" by over using the availability of hints, for example (Baker et al. 2006).

Serious games have been proffered as one solution to help maintain students' motivation, interest and engagement as they progress through digital learning environments (Adams and Clark 2014; Gee 2009; Jackson and McNamara 2013; Johnson and Valente 2008; McNamara et al. 2010; McQuiggan et al. 2010; Shaffer 2007; Sabourin et al. 2013). Serious games are games that teach, and they are usually digital-based. Gee (2009, 2013) has argued that good video games contribute to learning by affording experiences in good problem solving (e.g., focus on problems, provide clear goals, give feedback, create engagement through narratives, provide social interactions). There are several different types of game platforms crossing a spectrum from immersive 3D worlds to virtually embedded power-point presentations for individual players and multi-party games (Graesser et al. 2016). Across this large spectrum of types of serious games, common features exist including freedom on the part of the player, entertaining features, leveling up, and self-regulatory practices (Lepper and Henderlong 2000). These features manifest in many ways including fantasy, storylines, points, challenge, and competition (Ritterfeld et al. 2009). The reasoning for including these features is that they are believed to be motivating to the student player and thus increase persistence during game-play (Jackson and McNamara 2013; Landers 2014; Shaffer 2007). The increased engagement often corresponds to increased learning (McQuiggan et al. 2008; Rowe et al. 2011) through mediating factors, such as time on task (Landers and Landers 2014). On the other hand, there is evidence that serious games do not increase motivation. Wouters et al. (2013) conducted a meta-analysis of peer-reviewed research on serious games and learning. The results revealed that although serious games increase learning, the games did not significantly increase motivation. The authors speculate that one reason for this finding is that intrinsic motivation may be curtailed when serious games are chosen by the instructor rather than by the individual, as choice is an important feature of motivation (Shaffer 2007; Wouters et al. 2013).

1.3 AutoTutor

The purpose of the present study was to examine whether implementing game-based features to an existing ITS would affect learning and to a lesser degree, motivational and other noncognitive components. The ITS that we adopted is called *AutoTutor*. We chose AutoTutor for two reasons. First, it contains features common to ITSs (e.g., providing content, problems, feedback, adapting to the student's level of performance), and therefore,

results should be able to generalize to other ITSs. Second, because one of the authors had created the ITS (Art Graesser), we had the software to revise in order to create new conditions to address the research questions.

AutoTutor is a dialogue-based ITS that has natural language conversations between an animated pedagogical agent (or agents) and the human student (Graesser 2016; Graesser et al. 2004), which teaches topics such as computer literacy, comprehension, electronics, and physics. AutoTutor teaches students by simulating the dialogue moves of human tutors (Graesser et al. 2012b; Graesser and Person 1994; Graesser et al. 1995). AutoTutor's dialogues are organized around problems that require reasoning and explanations. The primary method of scaffolding good student answers is through *expectation and misconception-tailored (EMT) dialogue*. Both AutoTutor and human tutors (Graesser and Person 1994) typically have a list of anticipated good answers (called *expectations,* e.g., force equals mass times acceleration) and a list of anticipated *misconceptions* associated with each main question. AutoTutor guides the student in articulating the expectations through a number of dialogue moves: *pumps* (what else?), *hints*, and *prompts* for specific information. As the learner expresses information over many turns, the list of expectations is eventually covered, and the main question is scored as answered.

The EMT dialogue within AutoTutor is quite successful as there is substantial evidence accumulated over decades supporting the fact that students learn from AutoTutor (Graesser et al. 2016; Nye et al. 2014). Graesser and colleagues report that on the basis of several experiments comparing AutoTutor to different control conditions (do nothing, read transcripts of AutoTutor, read textbook), there is an average effect size of .80 (Graesser et al. 2004, 2012a, b; VanLehn et al. 2007).

In its original form, there are few if any aspects of AutoTutor that could be considered game-like. There are no stories, competition, points or other features that are common to games. However, there are versions of AutoTutor that include simulations, such as a visual interface that can be manipulated by the user (Graesser et al. 2005). Simulations may be considered somewhat game-like in that the user controls a virtual system, noting how the system changes according to the user input. This is similar to games in that the current state of the program or interface changes according to the actions of the user. However, these simulations did not involve adding a narrative or other game-like features. More recently, a game-like version of AutoTutor (called *Operation ARIES*) was created to teach research methodology, as will be discussed in the Methods section.

1.4 The Effects of Game Components in Intelligent Tutoring Systems

Unfortunately, there is very little research that has investigated how game-based features would affect learning and motivational components when added to an already established ITS (Graesser et al. 2016). Jackson and McNamara (2013) have conducted some research on this topic by examining two versions of an ITS called iSTART that teaches self-explanation, a successful strategy in reading, to high-school students (McNamara 2004). In that study, they compared the traditional version of the ITS with a game-based version of the ITS called iSTART-ME (Motivationally Enhanced) using high school students. In iSTART, the concept of self-explanation is first defined and illustrated via animated agents, and the user is then able to practice providing self-explanations with feedback given by the animated agents. In iSTART-ME, users are able to accumulate points to spend through increased performance over time. Players advance through levels, unlock new features, play mini-games, and have the opportunity to personalize a character. A comparison of the two implementations revealed that both versions produced similar learning gains across

eight sessions, but the game-based version maintained motivation across the session, whereas motivation showed a slight decrease across the same time period in the ITS control condition (Jackson and McNamara 2013). The findings also suggested that enjoyment increased across time for the game-based version, whereas enjoyment decreased and then leveled out for the ITS version.

In contrast to Jackson and McNamara (2013), which showed that the game-based features did not affect learning for high school students, other research suggests that narratives, a common feature of games (but not included in iSTART-ME), may *decrease* learning for college students. Adams et al. (2012) tested whether the presence of a narrative within two narrative discovery games (i.e. *Crystal Island*; Rowe et al. 2011) and *Cache 17* (Koening 2008) impacted learning in college students. In Crystal Island, the player moves about in a 3D environment on a remote island while interviewing inhabitants in an attempt to identify the source of a disease, a task common in 8th grade microbiology. In Cache 17, the player learns about electrical circuits and electromechanical devices while interacting in a 3D narrative world where he or she must open locks to find stolen paintings from World War II. In the first experiment, Adams et al. (2012) investigated the impact of the narrative on learning by comparing the full Crystal Island game to a non-narrative version of the game that was presented in slide-show format. Results suggested significantly greater retention and transfer in the slideshow (non-narrative) condition. However, it was difficult to isolate the cause of the finding because the conditions differed on both media (slideshow vs. game) and on narrativity (absent vs. present). In a follow-up experiment using Cache 17, narrative and non-narrative versions of the game were compared to a slide-show version. The narrative and non-narrative versions differed in that in the latter, there was no introductory video that explained the story context to the player. They found higher learning gains for the slideshow condition and equal gains for the narrative and non-narrative conditions. In summary, the findings of the available studies contradict the notion that hands-on activities within a scenario improve learning (i.e., what they refer to as the Discovery Hypothesis) and instead, a more static slideshow presentation was superior, a finding which was based on two games.

Although Adams et al. (2012) failed to find a difference between a narrative and non-narrative version of Cache 17, a study by McQuiggan et al. (2008) using 8th grade middle school students did find a difference between versions of Crystal Island which differed only on the extent or dosage of narrativity. In that study, participants interacted with one of three conditions of Crystal Island: high-narrativity, low-narrativity, and powerpoint (slideshow). In the high narrativity condition, participants interacted with the default version of Crystal Island (including the narrative about research members) with extensive back-stories about characters falling ill and an especially interesting scenario where members have been poisoned. In the low-narrativity condition, the poisoning storyline and back-stories about the characters were excluded, leaving only the overall setting (people are getting ill) intact. In the powerpoint condition, slides from the game were used without any storyline elements. Participants also took several measures including presence, interest, achievement goals, and science self-efficacy. Regarding learning gains, they report the following pattern of significant differences: powerpoint > low-narrativity > high narrativity. In addition, the high narrativity condition led to greater presence scores than the low narrativity condition. Therefore, learning was increased when narrative elements were removed from the learning environment, a finding largely consistent with Adams et al. (2012).

Although narrativity is only one aspect of games, the studies by Adams et al. (2012) and McQuiggan et al. (2008) do suggest that adding a storyline may have either no effect or be detrimental to learning. The interpretation of why this occurs is open to speculation, but the

authors posit that understanding the story consumes cognitive resources that otherwise could be allocated to the primary learning objectives, a theory aligned with *cognitive load theory* (Sweller 1988, 1999). The basic premise of cognitive load theory is that there is a limited amount of mental resources available at any given time. Specifically, the theory assumes that there are three types of cognitive load or effort exerted in working memory as a person interacts with instructional interfaces: intrinsic, germane, and extraneous. Intrinsic load refers to the resources needed to understand the elements necessary for comprehending the targeted topic of learning (e.g., microbiology). Germane load refers to the effort required for linking a current topic with prior knowledge to create a schema of the new topic. Extraneous load refers to the effort needed to process unnecessary information, such as the design, look, and navigation of the instructional materials. According to this account, the presence of a story causes extraneous load because it is unnecessary for learning the targeted topic. Hence, the story causes the learner to use resources that otherwise would be allocated to intrinsic and germane load necessary for creating a deep mental model of the material. The story is essentially a distraction. However, more research is necessary to make this important conclusion.

It is unclear how adding game-like components to AutoTutor will affect learning and noncognitive states like enjoyment and engagement, if they do at all. The uncertainty arises because previous research gives contradictory evidence in regards to these two constructs (i.e., learning and noncognitive states) in environments with components that are quite different from those added to the game-like version of AutoTutor. On the one hand, the results of Jackson and McNamara (2013) would indicate that the game-like features would lead to higher positive effects on noncognitive variables (e.g. motivation) than a non-game control, but not on measures of learning, which would remain constant across both conditions. However, these results may not apply here because the game-like features in the present experiment are very different from iSTART-ME (the gamified version of iSTART used by Jackson and McNamara 2013). Specifically, there is a narrative in the game-like version of AutoTutor, but there was no narrative in iSTART-ME. On the other hand, the presence of a narrative may decrease learning because of additional extraneous load (Adams et al. 2012; McQuiggan et al. 2008). However, those findings were based on games emphasizing discovery learning, where a learner is required to explore and discover implicit targeted learning objectives in a 3D environment, neither of which was a feature of the game-like version of AutoTutor.

Based on the three studies mentioned above, we would expect that the game-like version of AutoTutor would either have no effect (Jackson and McNamara 2013) or a detrimental effect on learning (Adams et al. 2012; McQuiggan et al. 2008). In regard to noncognitive factors, we might discover an increase in motivation, which would be consistent with Jackson and McNamara (2013) and McQuiggan et al. (2008). However, it is possible that we would find no difference based on the meta-analysis of Wouters et al. (2013).

2 Experiment 1

2.1 Methods

2.1.1 Participants

Fifty undergraduate psychology students (39 % male) enrolled in an undergraduate research methods course at Northern Illinois University served as participants in this

experiment. The approximate average age was 23. The prerequisites for enrolling in the course were having passed an introductory course in psychology and having earned a C or better in an introductory course in statistics. A primary reason for using students enrolled in a research methods course is that they would learn highly relevant information to the course from participating. Another reason is that the experiment was run over 4 1-h sessions which would not attract many participants without substantial inducements. The participants were given course credit for participating.

2.1.2 Design

The experiment used a pre-test, post-test randomized control design.

2.1.3 Materials

2.1.3.1 Game-Like Version of AutoTutor The game-like version of AutoTutor utilized one learning module out of three in a serious game called *Operation ARIES* (Acquiring Research Investigative and Evaluative Skills; Millis et al. 2014). Operation ARIES was created by the authors to help students learn scientific inquiry skills in the domains of psychology, biology and (to some extent) chemistry. As noted in the Introduction, students in the United States have poor knowledge regarding scientific inquiry skills (National Science Foundation 2012), thus we decided that these skills would be the primary focus of the game.

From conception, Operation ARIES included both an ITS (AutoTutor) and game-like components (e.g., points, competition and a narrative). However, the developers had never tested the effects of adding the game-like components on learning and noncognitive factors, hence the need for the current experiment. Also, it should be noted that the narrative was created to align with the pedagogical aspects of the game. In brief, the narrative involved extra-terrestrial creatures masquerading as human beings and circulating bad research. The human student is charged with learning research methodology to uncover the extra-terrestrials by identifying bad research. The storyline was presented across three teaching modules that teach factual, applied, and question generation about 12 topics of research.

Operation ARIES has three learning modules, but because of time limitations, we could only test one module. The module that we used is referred to as the "Case Studies" module. In this module, students must identify and name one or more flaws in research summaries that we refer to as research cases. The entire list of flaws were: poor or missing comparison group, no random assignment, dependent variable (DV) could be more sensitive, accurate, or precise, DV is not scored objectively, DV is not valid, subject bias, mortality or attrition, small sample size, poor sample selection, experimenter bias, premature generalization of results, and confuse correlation with causation. The player competes against another artificial agent for game points. There are two additional artificial agents: the teacher agent "Dr. Quinn" who helps students identify flaws in research cases with hints and prompts, and an alien agent, "Broth" who helps carry the storyline. The storyline is updated between research cases by snippets of conversation between the agents and email messages that appear on the screen. It is noteworthy to mention that there are additional pedagogical techniques in this game such as conversations and an available hint list of potential flaws as well as an E-text. However, from previous investigations, it appears the students rarely use the hint list or E-text in this module (Forsyth 2014).

Table 1 Sample research case

Let's dance
Stephanie Webber, a dance instructor at a local college, is always looking for new ways to help dance students improve their techniques. One way she thought to do this was to show a dancing video that she strongly believed would help improve dance ability. The video showed dance performances from the popular TV show "Dancing with the Stars"
Before using the video with her own class, she decided to test out its effectiveness with randomly chosen shoppers at a local mall. She solicited shoppers by asking if they would be interested in participating in a research study involving dancing abilities. Ten shoppers signed up
When her study began, she measured the ten participant's dancing abilities using a test in which she had them dance alone while holding a broom. All dancers were videotaped, and their movements were coded using an objective scoring technique that has been validated in a number of previous studies. Webber, who has been trained in the technique, did the scoring herself
The ten participants at the mall completed the broom test at their own pace and then watched the video. After they watched the video, they took the test again, and once again, they were videotaped, and the tapes were scored
Webber found that the participants' dancing abilities were significantly better after they watched the video than they were before. In order to be certain that the video was truly effective, she repeated the study with another 10 dance students and found the same result. Webber feels confident that watching the dancing video does improve dancing skills, and she plans to use the findings in an advertisement for her dance studio
Correct flaws:
No control/comparison group
Small sample size
Poor sample selection
Experimenter bias

2.1.3.2 Flaw-Identification Tests Two forms of a flaw identification test were created by the authors for measuring learning performance. The tests each contained three different research cases that contained between three to five flaws. Each of the 12 flaws from Operation ARIES was present in one of the cases (across each form). In addition, there was a research case in each form that fell under the content of psychology (e.g., a pill that improves memory), biology (e.g., second hand smoke and lung disease), and chemistry (the effects of a plastic on hormone levels). The research cases were written in the same "popular press" journalistic style as the cases that occurred in the experimental intervention. The tests were printed in booklet form, with each case appearing on its own page. The instructions were to read each research case and write down any flaws regarding the research design or interpretation of the findings. In Forms A and B, the average length of the research cases was 344 and 361 words, respectively. The average Flesch-Kincaid reading level was 10.8 and 11.4, respectively. An example research case is presented in Table 1.

2.1.3.3 Surveys Although the primary focus of the present experiments was to document the effects of game-like features on learning, we measured some noncognitive states as well. Of course, there are dozens, if not hundreds, of variables one could measure in regard to noncognitive states and user experiences. Because of time limitations, we wanted a measure that would be relatively quick to administer. Based on published work in serious games (Jackson and McNamara 2013; Wang et al. 2009; Wouters et al. 2013), we created a user experience survey that measured constructs that we thought intuitively might be affected by game components. The constructs were engagement (How engaged were

you?), enjoyment (How much did you enjoy what you were doing?), motivation (How motivated were you in answering correctly?), frustration (How frustrated were you?), interest (How interesting was this to you?), challenge (How challenging was it?), choice (How much choice did you have in what you were doing?), self-perception of learning (How much did you learn?), and whether the student would recommend the program to a friend (Would you recommend this as a homework activity?). All of the questions used a Likert-type of scale (1 = not at all/nothing, 6 = very much/a lot).

2.1.4 Procedure

Participants were assigned to one of three conditions: game, text-only, and control. The participants were first given an informed consent form, which explained that the purpose of the study was to examine how people learn from experiments summarized in the press. The informed consent was followed by the pretest. Participants were randomly assigned to form A or B as the pretest, and the other form was assigned to them as their posttest. Participants took approximately 20 min to complete the pretest. The experiment was run early in the semester so that the flaws to-be-detected in the experimental conditions and pre- and post-tests were not addressed in classroom assigned readings or classroom experiences up to, and during, the time the experiment was underway.

Because participants had not read about the flaws in class (nor were they interacting in conditions where the flaws were explained such as in the game or text-only conditions), it was necessary to give them a brief primer about the flaws. Otherwise, identifying flaws would have no or little meaning for them. Therefore, all participants were shown a video that defined each flaw as well as a sheet with brief definitions. Participants were given this information after being given the pre-test. Participants in the game condition were also shown a video depicting how to interact with the program.

2.1.4.1 Game Condition Participants in the game condition (N = 15) played the "Case Studies" module of Operation ARIES. Because this module occurs in between two other modules of Operation ARIES, and because the story spans all three modules, it was important that participants in this condition were familiar with the story which preceded the point in time in which they were exposed. Otherwise, they would not understand it. Therefore, a summary of the story was given to the participants before they started working on the computer.

The order of events in the presentation of a research case in the game condition was:

1. Story relevant information was given by an email sent to the participant, a pop-up window, or a brief dialogue among the pedagogical agents.
2. The research case was shown on the computer screen in a text window, and presumably read by the participant.
3. The participant was asked to type in a flaw in a text box (e.g., "not enough subjects").
4. The program matched the input to the list of flaws, and if a match was made, it was shown to the participant (e.g., "We think you meant: small sample size"). If there was no match, the participant was asked to try again.
5. Feedback: if the matched flaw was correct, Dr. Quinn, the teacher agent, gave positive feedback (e.g., "Correct"), points were added to the participant's cumulative total, and the participant was asked to identify another flaw, if there were any remaining. If the answer was incorrect, negative feedback was given (e.g., "No") and points were subtracted from the participant's score, and the turn to identify flaws was passed to the

student competitor artificial agent, Tracy. If the participant gave two consecutive incorrect answers, and there were more flaws that were not identified, then the program initiated an AutoTutor tutorial dialog in step 6 for each of the remaining cases.

6. Dr. Quinn gave a hint (e.g., "think about the participants in the study" hint for "small sample size") followed by an opportunity for the participant to answer. If the answer to the hint was incorrect, then Dr. Quinn gave a prompt (e.g., "when there are few participants, the sample is said to be what?" (answer: small). Correct answers to either the hint or prompt were followed by feedback and additional points. If the prompt was incorrect, then the turn passed to the other player who was given the opportunity to answer the prompt.

When all flaws were covered via steps 1–5 (or 1 through 6), then the summary of the flaws present in the research cases was given to the participant. Figure 1 presents an annotated screenshot of the game interface.

2.1.4.2 Text-Only Condition Participants in the text-only condition (N = 16) were exposed to almost the same information as in the game condition but without any game attributes. In this condition, there was no story, no pedagogical agents, no competition, and no graphic interface. They read each case on the computer, and after each one, they were instructed to write down any flaws that they noticed. They then were given the correct answers along with explanations of why the flaws were present and they were also instructed to compare their answers with the correct answers. They proceeded in this way

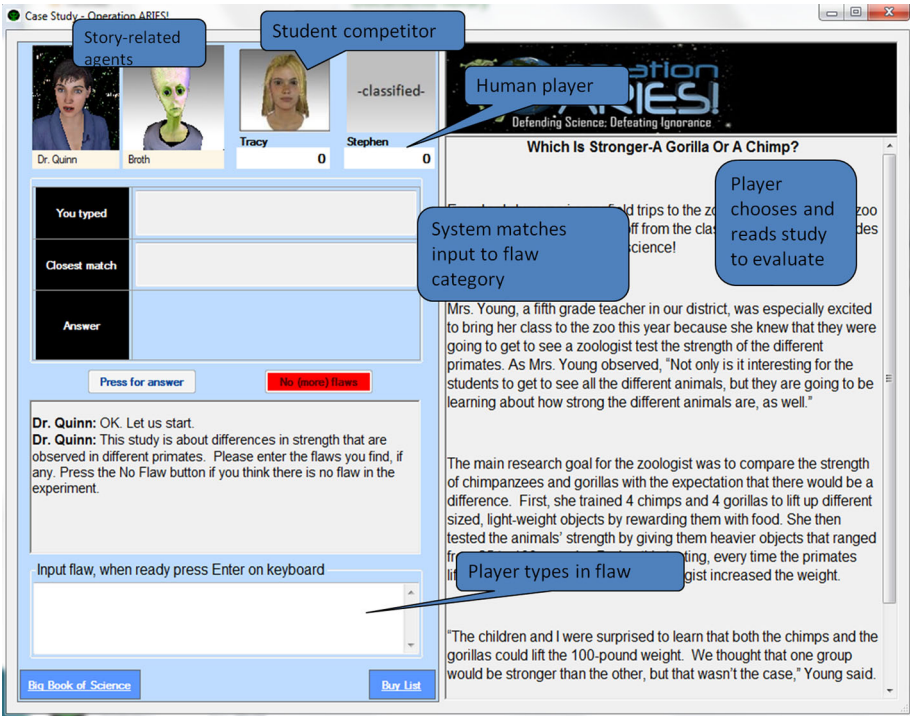

Fig. 1 Annotated screenshot of the game condition

until all cases were read and responded to. Therefore, the same cases and flaw summaries were present in both text-only and game conditions except that no tutorial dialogs occurred in the text-only condition. The text-only condition was presented using e-Prime, a computerized data collection software package. In both the game and text-only conditions, participants had a list of potential flaws available to them.

2.1.4.3 Control Condition Participants in the Control condition (N = 19) did not participate in the computer-based activities associated with the other two conditions. Instead, these students simply completed the pre- and post-test. They were given other course credit opportunities by participating in other research studies after the experiment was completed.

The experiment lasted 2 weeks, with a total of four 1-h sessions. Participants in the game and text-only conditions were given three cases on the first session, and five cases on each of the remaining three sessions. Sessions occurred either on Mondays and Wednesdays, or on Tuesdays and Thursdays. The same case studies were presented on the same day for both conditions. We should note that the multiple sessions were required to accommodate all of the cases. Based on pilot testing, we felt that the number of cases and sessions were needed to achieve significant learning gains.

The post-test was administered at the end of the last session. Following the post-test, the participants in the game and the text-only conditions completed the experience survey. To avoid interactions with course material, the experiment was run within the first month of class, and none of the material addressed in the experiment was included in the course readings, lectures or activities during the experiment. A flowchart summarizing the experimental procedure is shown in Fig. 2.

2.2 Results

2.2.1 Scoring of the Pre- and Post-Tests

Two raters scored each pre- and post-test. The raters were blind to condition. The raters determined which if any of the flaws from the list of 12 each participant listed for each summary. The raters were trained undergraduate students who had previously taken a research methods course. They were instructed to base their judgments on the meaning of what the participants wrote rather than just to identify key words. For example, if a participant wrote "needed more people in the study", then this was rated as "small sample size". Based on a sample of 20 participants' pre- and post-test tests, inter-rater reliability was acceptable, Kappa = .735.

We computed two "flaw identification scores" (FIS) for each participant, one for the pre-test and one for the post-test. For each, we first computed an overall hit rate. This was defined as the proportion of flaws present in the test cases that the participant correctly identified. Second, we computed the false alarm rate, which was defined as the proportion of flaws that were not present in the test cases that the participant wrote down. The FIS was computed by subtracting the false alarm rate from the hit rate. The FIS has a theoretical range from −1.0 to 1.0. A positive FIS would occur if the proportion of hits was larger than the proportion of false alarms, thereby indicating correct identification. A negative score would occur if the participant had a greater proportion of incorrect identification than correct identification. If a participant was guessing, the FIS would hover around zero.

We submitted the FIS to a 2 (time: pre vs. post) by 2 (pretest form: A vs. B) by 3 (condition: game, text-only, control) mixed ANOVA, with time as the within participant

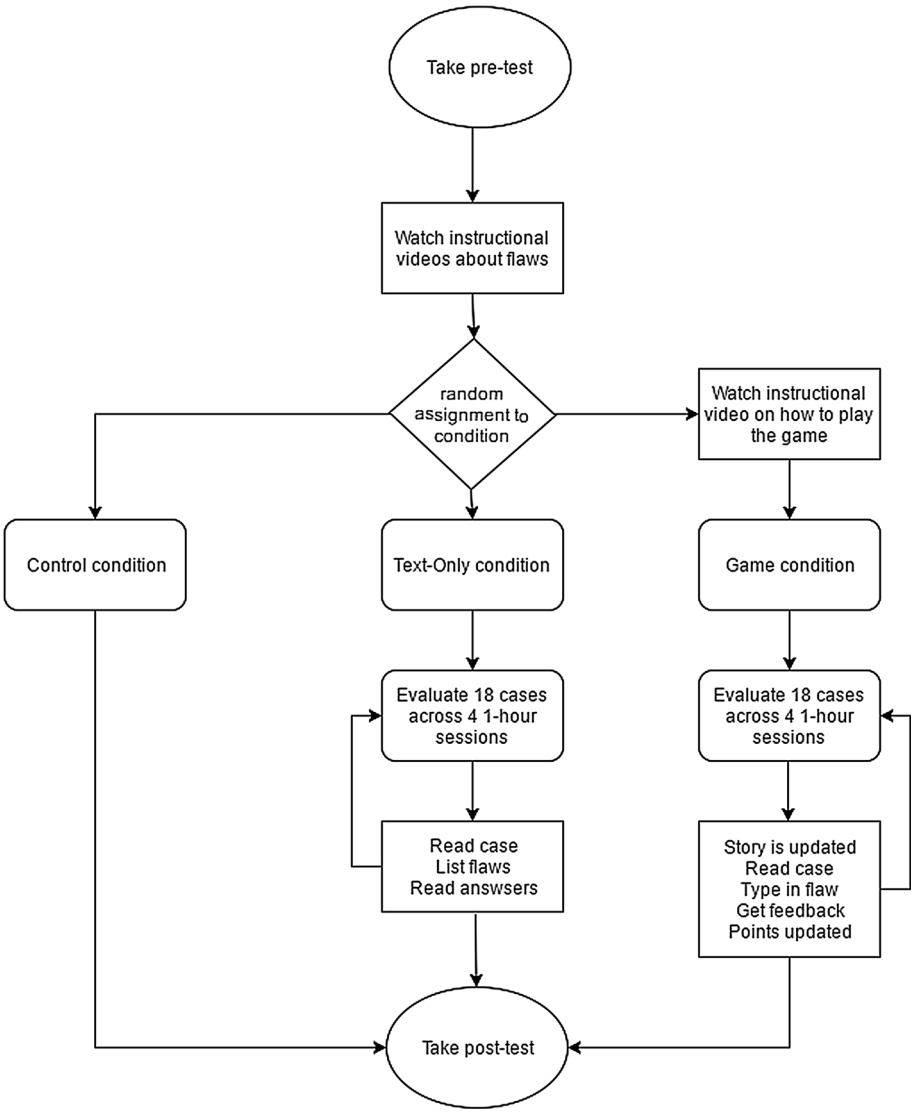

Fig. 2 A schematic of the experimental procedure

factor. The predicted time by condition interaction was significant, $F(2,44) = 36.85$, $p < .01$, MSe = .01, Partial Eta Squared = .63. The interaction is shown in Fig. 3.

As one can see, participants in each of the three conditions had very similar low pre-test scores (range .07–.09), whereas the post-test scores increased in both the game and text-only conditions (Ms = .43 and .32, respectively) but not in the control condition, which showed a slight decrease (M = .03). Therefore, students learned in both the game and text-only condition but there was no learning in the Control condition. Although the pre-test scores were low, the scores were significantly greater than zero (p's $< .05$). We tested the time by condition interaction without the control condition to address the question as to

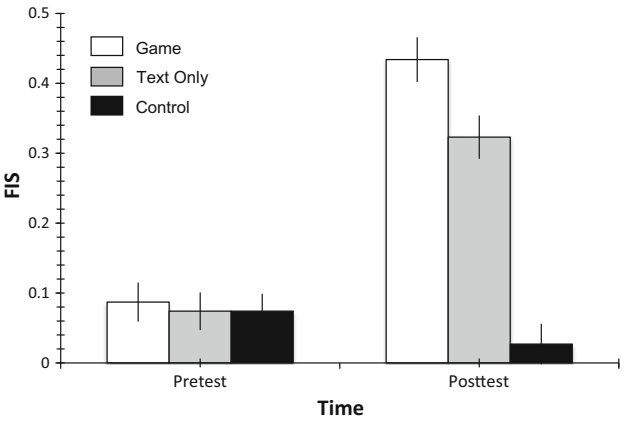

Fig. 3 Flaw identification scores (FIS) as a function of time of test (pre- vs. post-test) and condition

whether the increase from pre- to post-test was significantly greater in the game condition compared to the no game condition using the error term from the Omnibus F test. The interaction was marginally significant, $F(1, 44) = 3.60$, $p = .06$ (two-tailed). Therefore, it appears that the students in the game condition learned more than students in the text-only condition.

In addition to the condition by time interaction, the time by pretest form interaction was significant, $F(1,44) = 11.81$, $p < .01$. The difference between form A and form B was larger at pretest ($M = .12$ vs. .03) than at posttest ($M = .28$ vs. .24). The interaction suggests that without exposure to the treatment, Form A was easier than Form B, but with exposure to the treatment the difference was reduced. However, there was no main effect of pretest form ($p = .30$), no condition by pretest form interaction ($p = .23$), and no time by pretest form by condition interaction ($p = .11$). Therefore, pretest form did not appear to have moderated the condition by time interaction.

2.2.2 Survey Results

The mean responses to the survey questions are given in Table 2. The mean responses fell near the midpoint of the 6-point scale indicating moderate levels of each construct. A significant difference occurred between the game and text-only condition on frustration,

Table 2 Survey responses from Experiment 1

Construct	Game	Text-only
Engagement	3.56 (.96)	3.56 (.89)
Enjoyment	2.75 (1.12)	2.50 (.63)
Motivation	4.00 (.89)	3.68 (.70)
Frustration	2.87 (1.36)	2.00 (.96)**
Interest	3.37 (1.02)	2.81 (.83)*
Challenge	3.18 (.65)	3.18 (.83)
Choice	2.50 (.81)	2.62 (1.36)
Amount learned	4.00 (.51)	3.87 (.71)
Recommend	3.68 (1.13)	3.50 (1.41)

** $p < .05$ (two-tailed);
* $p < .05$ (one-tailed)

with the game version indicating more frustration ($M = 2.87$) than the text-only condition ($M = 2.00$), $t(30) = 2.09$, $p < .05$. It is difficult to gauge why participants in the game condition reported experiencing more frustration than participants in the text-only condition. One reason may have been that the game program matched the participant's typed flaw to a list of flaws using natural language processing algorithms. If the matching was incorrect even if the answer was correct, then it is likely that the participant would feel frustration. A significant difference occurred between the conditions on the level of interest, with participants in the game condition indicating slightly higher interest ($M = 3.37$) than participants in the text-only condition, $t\,(30) = 1.70$, $p < .05$ (one-tailed).

2.3 Discussion

The condition by time interaction suggests that the game condition marginally outperformed the nongame (text-only) condition. This finding was unexpected because previous research found that gamifying an ITS (Jackson and McNamara 2013) had little effect on learning and that the presence of a narrative had decreased learning (Adams et al. 2012; McQuiggan et al. 2008). Therefore, this finding by itself is somewhat noteworthy.

However, there are several limitations. There may have been too great of a difference between the game and text-only conditions to isolate the potential effects of the game-like components. The conditions were similar in that the same cases were analyzed by the participants on the same days, and participants received corrective feedback along with explanations why the flaws were present in each. However, beyond the presence of the storyline, points and competition, the conditions also differed on the complexity of the interface and the presence of tutorial dialogs in the game condition. Because we know that tutorial dialogs are effective for increasing learning, the question arises as to whether the greater performance in the game condition was due to the presence of AutoTutor dialogs or due to the game-like elements (competition, story, points).

The goal of Experiment 2 was to compare the game condition to a condition that did not have the game-like features but still contained the AutoTutor dialogs and other features of the interface. We refer to this latter condition as the no game condition. If the game condition outperforms the no game condition, then this would constitute strong evidence in favor of gaming features enhancing learning. Because the control condition showed no evidence of learning and the sample was virtually identical to Experiment 1, we omitted this condition from Experiment 2 with the expectation that a similar finding would occur in the absence of the intervention.

3 Experiment 2

3.1 Methods

3.1.1 Participants

Sixty-two undergraduate students enrolled in a research methods class at Northern Illinois University participated for course credit. The reason for the different numbers of students from Experiment 1 was that the experiments were done during different semesters (Fall vs. Spring) and that enrollments vary from one semester to the next. The proportion of females

and males, and the ages of the participants were approximately the same as in Experiment 1.

3.1.2 Design

The experiment used a pre-test, post-test randomized control design.

3.1.3 Procedure

Participants were randomly assigned to either the game condition or the no game condition. The materials and procedures were exactly like in Experiment 1 with two exceptions. The first was that there was no "do nothing" control condition in Experiment 2. We did not include a control condition because the primary reason for the experiment was to compare a game to a no game condition. The second was that in the no game condition, participants interacted with the Case Studies module of Operation ARIES but were not given the narrative about the aliens, updating of points, and competition. Tracy, the competitor in the game condition, also provided answers in the no game condition, but because there were no points accumulated or lost, we believed that participants would be less inclined to view Tracy as a competitor in the no game condition.

3.2 Results and Discussion

The FIS were computed in the same way as Experiment 1. The mean pre-test scores for the game and no name conditions were .10 (SD = .10) and .12 (SD = .15), respectively. The mean post-test scores were .31 (SD = .15) and .36 (SD = .13), respectively. Although the increase in scores from the pre- to the post-test was highly significant [$F(1, 59) = 100.55$, $p < .01$, MSerror = .01, partial eta squared = .63], no other effects were significant (all p's $> .27$). The time (pre- vs. post-test) by condition (game vs. no game) interaction, which would indicate higher learning gains in one condition over the other, was not significant, $F(1, 59) = .295$, $p < .60$, Partial Eta Squared = .005, observed power($1 - \beta_{\text{err prob}}$) = .08.

3.2.1 Survey Results

Overall, the means for the survey questions were similar in magnitude as those reported in Experiment 1. Unlike Experiment 1, however, no significant differences emerged between conditions in Experiment 2.

Although participants learned in both conditions, there were no differences between the game and no game conditions. The lack of a difference indicates that the game elements of points, storyline and competition together did not have an effect on learning. These findings imply that the difference between the game and text-only conditions on the FIS in Experiment 1 could be accounted for by having tutorial dialogs in the former condition. Therefore, it appears that interactivity via dialogs had greater impact on learning than game features, and that the game features did not significantly decrease learning.

4 General Discussion

The purpose of these studies was to address two research questions, the first of which was *can we build off of pre-existing work in Intelligent Tutoring Systems when creating serious games?* The answer appears to be yes with qualifications. On the one hand, one could add attributes found in video games and incorporate them into an ITS. Indeed, that is precisely how we decided to develop many aspects of Operation ARIES, which included adding a narrative, competition, and points to the existing framework of AutoTutor. A similar development strategy was largely implemented for iSTART-ME, a serious game created by Jackson and McNamara (2013), to "gamify" McNamara's original ITS called iStart. However, the purpose of adding game-like features to ITSs is to enhance motivation which then should increase behaviors (e.g., time-on-task) that enhance learning (Landers 2014). Under this assumption, the effectiveness of the game-like features may be measured by the amount that students' motivation is increased. We predicted that the game-like features would increase motivation similarly to Jackson and McNamara's findings (2012). Unfortunately, we found no significant difference between the game and nongame versions on motivation and on many of the other noncognitive features that we measured, Instead, the results support Wouters et al. (2013) which showed serious games (which may include some of the game-like features included here) may not increase motivation over other instructional control conditions. Therefore, one cannot just slap on points and a storyline to make an ITS more motivating. One silver lining here is that it appears that the game-like features did not reduce motivation.

In regard to additional noncognitive states, we found little difference between the conditions. However, we should mention that we only had one question (item) for each of the constructs included in the survey, including motivation. Typically, surveys and inventories require more than one item to measure a psychological construct. So, although we found that adding the game-like features did not reduce nor increase motivation, it is possible that with more sophisticated and psychometrically sound measuring instruments, there is a greater likelihood that an existing effect would be detected.

The second research question was *would adding game-like features to an existing ITS affect learning, and to a lesser extent, noncognitive states?* In Experiment 1, we found that the game-based ITS outperformed a "do nothing" control group and a text-based version of the ITS. This suggests that students were learning from the game-based scenario. In both the game-based and non-game conditions, participants read short summaries of research, wrote down design flaws, and then read the correct answers (feedback). The text-based version excluded the storyline, agents, competition and points, an interesting visual interface, and brief tutoring sessions when both the student and the virtual competitor failed to identify all flaws. Because tutoring was involved in the game-based version and not in the text-based version, the advantage for the game-based version might have been due to the tutorial dialogs. Therefore, Experiment 2 compared two versions that had very similar interfaces, but differed on only competition, points, and presence of a story line. The results indicated no differences on learning and motivation-related responses.

One answer to the second research question based on the present studies is that the presence of the storyline, competition and points may not significantly improve or detract from learning. The most direct comparison between the presence and absence of these features offered in Experiment 2 show no effect. The learning gains for the game and no game condition in that experiment were .21 and .24, respectively. The slight advantage for the no game condition is consistent with evidence that the presence of a storyline in a

discovery learning online context decreases learning (Adams et al. 2012; McQuiggan et al. 2008). However, the difference was small and the sample size was also small, resulting in low power (.08) indicating that we are less likely to detect the effect if it in fact exists in the present study. The lack of a significant effect for this particular interaction suggests to us that the presence of the game-like variables had minimal impact on learning.

It is interesting to ponder why other researchers have found stronger evidence that the storylines decrease learning (Adams et al. 2012; McQuiggan et al. 2008). This is a difficult question to answer definitively because there are many variables that vary across games, conditions/experiments, and platforms, making comparisons difficult (Gee 2009; Landers 2014). Even within a given game, testing individual features is difficult if not impossible because the features are intertwined to create the game-like atmosphere. To possibly aid in the understanding of how the present study fits in with the other relevant literature mentioned, we summarize features and major findings in Table 3.

We should mention that the studies summarized in Table 3 is an extremely small subset of studies that examined serious games and learning. They represent a select few that we had found that had either tried to measure the effects of "gamifying" an existing ITS (Jackson and McNamara 2013; the present study) on learning or had tried to examine the effect of narratives within a serious game on learning (Adams et al. 2012; McQuiggan et al. 2008). One factor that stands out has having a possible effect is the amount of time that participants interacted with the game. When participants interacted with the game for an hour, the narrative had appeared to decrease learning. In both cases when there were multiple sessions, there was no effect of the game-like variables on learning.

As been mentioned earlier, it is difficult to isolate factors across studies, and in Table 3, we see that clearly. The studies that differed on time on task also differed on the level of realism. Adams et al. (2012) and McQuiggan et al. (2008) examined serious games in which the player moves about within a 3D world (realistic), and these studies showed that the narrative had decreased learning. In contrast, the Operation ARIES interface is a schematic layout of textboxes and agent heads (Fig. 1), rather than a 3D world. (For simplicity, we refer to all other interfaces which depict a realistic looking virtual world as 'schematic'.) It is possible that the narrative, on top of a seductive 3D world, poses a sufficient amount of extraneous load that decreases learning of the targeted information; however, the narrative on top of a relatively static display may not. It is also possible that the storyline within Operation ARIES is not as engaging as the storyline in Crystal Island and Cache 17, the learning environments that show some detrimental effects of narrative worlds. Perhaps the more engaging the storyline is, the more attention is drawn away from the targeted learning activities, at least to the extent that the story is separated from the learning activities. Another difference lies within the cognitive resources needed to do the targeted learning activities. If the learning domain requires a lot of necessary cognitive resources, then storylines and other features of the environment may cause more harm than if the learning domain requires fewer resources. It is entirely possible that identifying flaws in research summaries required fewer cognitive resources than figuring out why people are getting sick on an island (Crystal Island) or understanding electronic locks (Cache 17).

Another variable to consider is the age of the participants. What may be an interesting storyline or a motivating factor to college students may not be interesting or motivating to younger students. McQuiggan et al. (2008) used 8th graders and found that narrativity decreased learning, whereas Adams et al. (2012) used college students and found similar results. However, like Adams et al. (2012), we also used college students and did not find a decrease on learning due to the game features (which included a storyline). Therefore, age does not appear to show a consistent effect across these studies. In fact, in their respective

Table 3 A summary of game characteristics and findings across five studies

Serious game	Domain	Publication	Grade level	Realism	Time on task	Competition/points	Story line	Major finding of game components
iSTART-ME	Reading strategies	Jackson and McNamara (2013)	College students	Schematic	8 1-h sessions	Points/trophies/ levels	No	Maintain motivation but does not increase learning
Crystal Island	Biology	Adams et al. (2012)	College students	Realistic	1-h	Trophies	Yes	Decrease learning
Crystal Island	Biology	McQuiggan et al. (2008)	8th graders	Realistic	1-h	Trophies	Yes	Decrease learning
Cache 17	Electronics	Adams et al. (2012)	College students	Realistic	1-h	Unknown[a]	Yes	Decrease learning
Operation ARIES	Research methods	Present study	College students	Schematic	4 1-h sessions	Yes	Yes	No effect on learning or motivation

[a] We do not know whether Cache 17 includes competition and/or points

meta-analyses, Vogel et al. (2006) and Wouters et al. (2013) reported no significant differences due to age. But, because of the inherent complexity of isolating factors that mediate game features and learning, such as age, more research is clearly warranted.

A related issue to learning within a narrative framework is whether the narrative world is aligned with the learning objectives. One could argue that the extent to which the story is not directly related to the learning objectives, the story may distract from learning those objectives. In these cases, the story becomes an example of stimuli that are referred to as "seductive details" which are interesting pieces of information in texts that are not important to understanding the phenomenon being described. For example, a seductive detail might be a picture of a lightning strike or a "fun fact" about lightning in a text about how lightning occurs. Seductive details are known to decrease learning (Harp and Mayer 1998), especially for individuals with low working memory capacities (Sanchez and Wiley 2006). From this perspective, many pieces of information in the story world in Crystal Island, Cache 17, and Operation ARIES would be considered seductive details regardless of the developer's intent to align the story elements to the material. What may serve as seductive details may be necessary to either maintain the storyline or the game atmosphere. However, this observation alone would not explain why we found little impact of the narrative on learning, whereas others (Adams et al. 2012; McQuiggan et al. 2008) have reported significant effects.

Lastly, there are two other possible reasons why we found a smaller effect than other researchers when adding a narrative. One is that in this study, participants interacted with the game (or nongame) across four sessions, whereas participants in Adams et al. (2012) and McQuiggan et al. (2008) only interacted with the game for around an hour. It is conceivable that proportionally more cognitive resources must be allocated to the narrative aspect when the game play is short. Over extended play, the story elements might be consolidated in long-term memory, and therefore, might have smaller load on working memory as the participant processes the to-be-learned information. Secondly, besides the presence of the storyline, in the current study, competition and points were manipulated whereas these features were not manipulated in Adams et al. (2012) or McQuiggan et al. (2008). We originally thought that all three of these features would have similar effects, but perhaps they had differing effects. It is possible that the story may have, in fact, decreased learning by imposing a distraction or extraneous load, but the competition may have increased learning by increasing motivation. If this was the case, then both of these could have cancelled each other out, leading to a zero sum gain.

The question of how game-like elements contribute or detract from learning is an important one because most everyone (students, educators, designers of educational software) want students to be engaged in the learning activities. No one wants students to be frustrated or bored because these emotional states often lead to disengagement (Graesser and D'Mello 2012). Emotional states are quite intricate in relation to learning. For example, there can be a fine line between frustration and confusion, the former being shown to correlate negatively with learning and the latter showing a positive correlation (D'Mello and Graesser 2012; Graesser et al. 2008). Confusion predicts learning when the learner can resolve the source of the confusion. Frustration occurs when there is no resolution. In order for designers of educational software to maximize learning, they need to understand the interactions among design features (e.g., competition, narratives), noncognitive states (e.g., emotions, motivation), and aspects of the learner (e.g., prior knowledge, interest) (Landers 2014). Obviously, this is a tall order, and future research is needed to further isolate the effects of game-like features on learning and non-cognitive states.

In sum, the present research suggests that we accomplished our goal of creating a game from an existing ITS. The interactive game accounted for higher learning gains than a control without similar game-like features. However, there was little difference between learning in the game versus no game environments. Therefore, our findings suggest that the narratives and game features may not decrease learning when added to an ITS. Because there are many design features to ITSs (including social factors which we did not address), it will be necessary to further track the alignment of learning, noncognitive states, and design features before we are certain how design features contribute to increased learning and persistence in advanced digitally-based learning environments.

Acknowledgments 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.

References

Adams, D. M., & Clark, D. B. (2014). Integrating self-explanation functionality into a complex game environment: Keeping gaming in motion. *Computers & Education, 73*, 149–159.

Adams, D. M., Mayer, R. E., MacNamara, A., Koenig, A., & Wainess, R. (2012). Narrative games for learning: Testing the discovery and narrative hypotheses. *Journal of Educational Psychology, 104*, 235–249.

Baker, R. S. J., Corbett, A. T., Koedinger, K. R., Evenson, S., Roll, Il, Wagner, A. Z., et al. (2006). Adapting to when students game an Intelligent Tutoring System. In M. Ikeda, K. D. Ashley, & T. W. Chan (Eds.), *Proceedings of Intelligent Tutoring Systems 8th international conference ITS 2011, May 2011* (pp. 392–401). Berlin: Springer.

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

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*, 123–152.

Blizzard Entertainment. (2004). *World of warcraft*. http://www.google.com/?gws_rd=ssl#q=world+of+warcraft&stick=H4sIAAAAAAAAAONgFuLQz9U3MMyqsFTiBLGMDFPKKrREwjJTUvPdE3NTfROzU4tCwEwAMaI3Ni0AAAA.

ChanMin, K., & Pekrun, R. (2013). Emotions and motivation in learning and performance. *Handbook of research on educational communications and technology* (pp. 65–75). New York: Springer.

D'Mello, S., & Graesser, A. C. (2012). Emotions during learning with AutoTutor. In P. J. Durlach & A. Lesgold (Eds.), *Adaptive technologies for training and education*. Cambridge: Cambridge University Press.

Dzikovska, M. O., Steinhauser, N., Farrow, E., Moore, J. D., & Campbell, G. E. (2014). BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education, 24*, 284–332.

Forsyth. (2014). *Predicting learning: A fine-grained analysis of learning in a serious game*. Unpublished doctoral dissertation. The University of Memphis.

Gee, J. P. (2009). Deep learning properties of good video games: How far can they go? In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 67–82). New York: Routledge.

Gee, J. P. (2013). Games for learning. *Educational Horizons, 91*, 17–20.

Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education, 26*, 124–132.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An Intelligent Tutoring System with mixed-initiative dialogue. *IEEE Transactions in Education, 48*, 612–618.

Graesser, A. C., Conley, M. W., & Olney, A. (2012a). Intelligent Tutoring Systems. In S. Graham & K. Harris (Eds.), *APA handbook of educational psychology*. Washington, DC: American Psychological Association.

Graesser, A. C., & D'Mello, S. (2012). Emotions during the learning of difficult material. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 57, pp. 183–225). Amsterdam: Elsevier.

Graesser, A. C., D'Mello, S. K., Hu, X., Cai, Z., Olney, A., & Morgan, B. (2012b). AutoTutor. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation and resolution* (pp. 169–187). Hershey, PA: IGI Global.

Graesser, A. C., D'Mello, S. K., Craig, S. D., Witherspoon, A. M., Sullins, J., McDaniel, B., et al. (2008). The relationship between affective states and dialogue patterns during interactions with AutoTutor. *Journal of Interactive Learning Research, 19*(2), 293–312.

Graesser, A. C., Hu, X., Nye, B., & Sottilare, R. (2016). Intelligent Tutoring Systems, serious games, and the Generalized Intelligent Framework for Tutoring (GIFT). In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulation for teaching and assessment* (pp. 58–79). Abingdon: Routledge.

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science, 23*, 374–380.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A. M., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral, Research Methods, Instruction & Computers, 36*, 180–193.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*, 104–137.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*, 495–522.

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

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology, 90*, 414–434.

Institute for Education Sciences. (2009). *Science 2009: National assessment of educational progress at grades 4, 8, and 12*. http://nces.ed.gov/nationsreportcard/pdf/main2009/2011451.pdf.

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based Intelligent Tutoring System. *Journal of Educational Psychology, 105*, 1036–1049.

Johnson, L. W., & Valente, A. (2008). Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. In M. Goker & K. Haigh (Eds.), *Proceedings of the twentieth conference on innovative applications of artificial intelligence* (pp. 1632–1639). Menlo Park, CA: AAAI Press.

Koening, A. D. (2008). *Exploring effective educational video game design: The interplay between narrative and game-schema construction* (Unpublished doctoral dissertation). Arizona State University.

Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & Gaming, 45*(6), 752–768.

Landers, R. N., & Landers, A. K. (2014). An empirical test of the theory of gamified learning: The effect of leaderboards on time-on-task and academic performance. *Simulation & Gaming, 45*(6), 769–785.

Lane, H. C., Noren, D., Auerbach, D., Birch, M., & Swartout, W. (2011). Intelligent tutoring goes to the museum in the big city: A pedagogical agent for informal science education. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education: 15th International conference* (pp. 155–162). Heidelberg: Springer.

Lepper, M. R., & Henderlong, J. (2000). Turning "play" into "work" and "work" into "play": 25 years of research on intrinsic versus extrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 257–307). San Diego, CA: Academic Press.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*, 1–30.

McNamara, D. S., Jackson, G. T., & Graesser, A. C. (2010). Intelligent tutoring and games (ITaG). In Y. K. Baek (Ed.), *Gaming for classroom-based learning: Digital role-playing as a motivator of study* (pp. 44–65). Hershey, PA: IGI Global.

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

McQuiggan, S. W., Rowe, J. P., Lee, S., & Lester, J. C. (2008). Story-based learning: The impact of narrative on learning experiences and outcomes. *Intelligent Tutoring Systems: Lecture Notes in Computer Science, 5091*, 530–539.

Millis, K., Graesser, A., & Halpern, D. (2014). Operation ARA: A serious game that combines intelligent tutoring and learning principles to teach science. In V. Benassi, C. E. Overson, & C. M. Hakala, (Eds.) *Applying the science of learning in education: Infusing psychological science into the curriculum*. Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/asle2014/index.php.

National Science Foundation. (2012). *National Science Board's Science and Engineering Indicators 2012*. http://www.nsf.gov/statistics/seind12/c7/c7h.htm.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education, 24*, 427–469.

Olney, A., D'Mello, S. K., Person, N., Cade, W., Hays, P., Williams, C., et al. (2012). Guru: A computer tutor that models expert human tutors. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of Intelligent Tutoring Systems (ITS) 2012* (pp. 256–261). Berlin: Springer.

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.

Richards, J., Stebbins, L., & Moellering, K. (2013). *Games for a digital age: K-12 market map and investment analysis*. New York: The Joan Ganz Cooney Center at Sesame Workshop.

Ritterfeld, U., Cody, M., & Vorderer, P. (Eds.). (2009). *Serious games: Mechanisms and effects*. New York and London: Routledge, Taylor & Francis.

Rowe, J., Shores, L. R., Mott, B., & Lester, J. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, *121*, 115–133.

Rus, V., D'Mello, S., Hu, X., & Graesser, A. C. (2013). Recent advances in intelligent systems with conversational dialogue. *AI Magazine, 34*, 42–54.

Sabourin, J. L., Rowe, J. P., Mott, B. W., & Lester, J. C. (2013). Considering alternate futures to classify off-task behavior as emotion self-regulation: A supervised learning approach. *Journal of Educational Data Mining, 5*, 9–38.

Sanchez, C. A., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition, 34*, 344–355.

Shaffer, D. W. (2007). *How computer games help children learn*. New York, NY: Palgrave.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.

Sweller, J. (1999). *Instructional design in technical Areas*. Camberwell, VIC: Australian Council for Educational Research.

VanLehn, K. (2011). The relative effectiveness of human tutoring, Intelligent Tutoring Systems, and other tutoring systems. *Educational Psychologist, 46*, 197–221.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, P. A. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 30*, 1–60.

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34*, 229–243.

Vygotsky, L. S. (1978). *Mind in society: Development of higher psychological processes* (p. 86). Cambridge: Harvard College.

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

Woolf, B. P. (2009). *Building Intelligent Tutoring Systems*. Burlington, MA: Morgan Kaufman.

Wouters, P., van Nimwegen, C., van der Spek, E. D., & van Oostendorp, H. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*, 249–265.