The impact of individual differences on learning with an educational game and a traditional ITS

G. Tanner Jackson*
Learning Sciences Institute,
Arizona State University,
Payne Hall, Tempe, AZ 85287, USA
E-mail: TannerJackson@asu.edu
and
Educational Testing Service,
660 Rosedale Rd., MS 16-R,
Princeton, NJ 08541, USA
E-mail: gtannerajckson@gmail.com
*Corresponding author

Laura K. Varner
Learning Sciences Institute,
Arizona State University,
Payne Hall, Tempe, AZ 85287, USA
E-mail: Laura.Varner@asu.edu

Chutima Boonthum-Denecke
Department of Computer Science,
Hampton University,
100 E. Queen Street, Hampton, VA 23668, USA
E-mail: chutima.boonthum@hamptonu.edu

Danielle S. McNamara
Learning Sciences Institute,
Arizona State University,
Payne Hall, Tempe, AZ 85287, USA
E-mail: Danielle.McNamara@asu.edu

Published in October, 2013

The impact of individual differences on learning with an educational game and a traditional ITS

G. Tanner Jackson*

Learning Sciences Institute,
Arizona State University,
Payne Hall, Tempe, AZ 85287, USA
E-mail: TannerJackson@asu.edu
and
Educational Testing Service,
660 Rosedale Rd., MS 16-R,
Princeton, NJ 08541, USA
E-mail: gtannerajckson@gmail.com
*Corresponding author

Laura K. Varner

Learning Sciences Institute,
Arizona State University,
Payne Hall, Tempe, AZ 85287, USA
E-mail: Laura.Varner@asu.edu

Chutima Boonthum-Denecke

Department of Computer Science,
Hampton University,
100 E. Queen Street, Hampton, VA 23668, USA
E-mail: chutima.boonthum@hamptonu.edu

Danielle S. McNamara

Learning Sciences Institute,
Arizona State University,
Payne Hall, Tempe, AZ 85287, USA
E-mail: Danielle.McNamara@asu.edu

Abstract: Educational games have the potential to provide motivating, effective training; however, the efficacy of these systems is unclear, and evaluations often fail to identify the relative impact of individual differences on learning outcomes. The current study aims to address these issues by comparing the learning gains from an educational game (iSTART-ME) and an intelligent tutoring system (iSTART). High-school students (n = 125) received comprehension strategy training from the two systems, and results indicated that both training environments yielded significantly better scores on posttest performance and learning measures than students assigned to a time-delayed control condition. Additionally, for both training conditions, students with a

low prior 'commitment to reading' exhibited the highest performance improvements. Overall, results indicate that educational games can produce learning equivalent to intelligent tutoring systems, and that this training can provide a means to overcome initial deficits for students with a low 'commitment to reading'.

Keywords: intelligent tutoring systems; ITSs; learning; educational games; reading comprehension; strategy training; individual differences.

Reference to this paper should be made as follows: Jackson, G.T., Varner, L.K., Boonthum-Denecke, C. and McNamara, D.S. (2013) 'The impact of individual differences on learning with an educational game and a traditional ITS', *Int. J. Learning Technology*, Vol. 8, No. 4, pp.315–336.

Biographical notes: G. Tanner Jackson is an Assistant Research Professor in the Learning Sciences Institute at Arizona State University. His research focuses on the design, development, and evaluation of game-based educational environments and adaptive learning systems that incorporate natural language processing. He is interested in how elements of these environments (both individually and in combination) affect users' cognitive and motivational constructs to produce an effective and enjoyable experience. Furthermore, since completing this article, he has started a position as a Research Scientist at the Educational Testing Service.

Laura K. Varner is a PhD student in Cognitive Science and the Learning Sciences Institute at Arizona State University. Her research primarily examines the cognitive processes, attitudes, and abilities that underlie proficiency across both text comprehension and production. She uses a variety of techniques, such as natural language processing, to capture the ways in which individual differences manifest in students' behaviours and is also interested in the impact of these factors on second language learning.

Chutima Boonthum-Denecke is an Associate Professor in the Department of Computer Science, School of Science, at Hampton University. She earned her PhD in Computer Science from Old Dominion in 2007; MS in Applied Computer Science from Illinois State University in 2000; and BS in Computer Science from Srinakharinwirot University in 1997. Her research interests include artificial intelligence (natural language processing, computational linguistics), intelligence tutoring system, information retrieval, web development technology, cognitive robotics, and integrating AI in cybersecurity fields.

Danielle S. McNamara is a Professor in the Psychology Department and Senior Scientist in the Learning Sciences Institute at Arizona State University. Her academic background includes a Linguistics BA (1982), a Clinical Psychology MS (1989), and a Cognitive Psychology PhD (1992). She develops educational technologies and conducts research to better understand cognitive processes involved in comprehension, knowledge and skill acquisition, comprehension strategies, and writing. Her research also involves the development and assessment of natural language processing (e.g., Coh-Metrix), game-based, intelligent tutoring systems (e.g., iSTART, Writing Pal; see http://www.soletlab.com), and the use of interactive dialog in automated tutoring systems.

This paper is a revised and expanded version of a paper entitled 'A comparison of gains between educational games and a traditional ITS' presented at the 25th Annual FLAIRS Conference, Marco Island, FL, USA, May 2012.

1 Introduction

Intelligent tutoring systems (ITSs) offer adaptive, potentially individualised training and practice on a wide variety of skills and content domains. These systems also automate certain aspects of instruction, thereby increasing opportunities for tutoring and learning without burdening teachers. Their success in producing consistent learning gains has been well documented. Indeed, ITSs have been found to be as effective as human instructors in terms of developing students' content knowledge and comprehension skills (VanLehn, 2011).

To achieve these learning gains, ITSs that focus on the acquisition and refinement of skills and strategies [e.g., self-explanation (SE) and metacomprehension] tend to require a commitment to practice and application over long time periods (Newell and Rosenbloom, 1981). Unfortunately, over time, students can lose focus and disengage from the learning environments (Bell and McNamara, 2007; D'Mello et al., 2007; Jackson and McNamara, in press). When students are not engaged, they are more likely to be bored or inattentive, neither being conducive to learning (e.g., Craig et al., 2004). Bored learners are more likely to bypass the system (Rodrigo et al., 2007) and less likely to actively reengage in constructive learning processes (Boekaerts et al., 2000; D'Mello and Graesser, 2006; D'Mello et al., 2007).

One method for improving engagement has been to incorporate game-like components into educational environments (for a review, see Clark et al., 2009). Well-designed games are appealing, partially because they address affective states, motivation, and expectancies of the player (O'Neil et al., 2005). A general assumption among educational game researchers is that games can improve students' motivation and engagement (among other things) and, consequently, enhance learning outcomes. Beyond individual studies that show positive motivational and learning outcomes (e.g., Ricci et al., 1996; Rowe et al., 2011), meta-analyses have reported that across groups of people (e.g., gender, age), interactions with games can lead to better outcomes for cognition, increases in skill mastery, and improved affect (Vogel et al., 2006; Wilson et al., 2009).

One overarching benefit of educational games is that they function similarly to sophisticated tutoring systems by providing the opportunity for adaptive, individualised learning. They afford a means for individualised practice with content and skills wherein instructors are potentially able to monitor the progress of the learners. Additionally, the rapid feedback within educational games can help learners to better regulate their progress and activities. Indeed, the feedback in a number of learning environments has been shown to significantly improve engagement (Anderson et al., 1995; Corbett and Anderson, 1990; Foltz et al., 2000). Finally, to a greater extent than traditional tutoring systems, games should help to render practice more enjoyable for learners, thus leading to perseverance and increased motivation to engage with systems across extended training sessions.

In addition to bolstering enjoyment and engagement, games can fulfil a number of educational purposes (Gredler, 2004). For instance, games can be used for assessment, in which the game evaluates whether the learner can accurately apply skills and knowledge. They can also be used to aid the acquisition of new knowledge, as they provide a practice environment that allows students to repeatedly apply relevant knowledge and skills in a variety of contexts (e.g., Orbach, 1979; Shank and Neaman, 2001). Additionally, educational games can help learners refine and integrate existing knowledge to develop a

stronger understanding of conceptual relationships; this allows students to create novel combinations of existing knowledge (e.g., Swaak and de Jong, 2001).

Some research suggests that game-based environments may interfere with initial learning, however, despite temporary decreases in immediate performance, game-based environments have shown learning gains comparable to ITSs (Jackson and McNamara, in press). These comparable results are achievable by synthesising the affordances of effective game design with powerful ITS learning principles, thus creating an environment with the power to promote and sustain motivation, engagement, and persistence and, as a result, increase learning. One example of this pursuit is the interactive strategy training for active reading and thinking-motivationally enhanced (iSTART-ME) tutor, which was built on top of an existing ITS (called iSTART) and adapted into a game-based environment. In this system, students can practice strategies, earn points, advance through levels, purchase rewards, create a personalised avatar, and play educational mini-games. The remainder of this work describes the two iSTART systems and discusses an experimental comparison of the two environments.

2 iSTART

iSTART is an ITS designed to improve students' reading comprehension by teaching SE in combination with effective reading strategies. iSTART introduces students to the concept of SE (i.e., explaining a text, or part of a text, to oneself) and provides instruction on how to use reading comprehension strategies to improve their understanding of difficult science texts. The development of iSTART was based on previous research with a successful human intervention called self-explanation reading training (SERT: McNamara, 2004; O'Reilly et al., 2006). This training was designed to help those low ability students who might not effectively use strategies on their own. Students who have been provided with iSTART have shown significant improvement in reading comprehension, comparable to the performance within SERT (Magliano et al., 2005). iSTART training is separated into three distinct modules that instantiate the pedagogical principle of modelling-scaffolding-fading: introduction, demonstration, and practice, respectively.

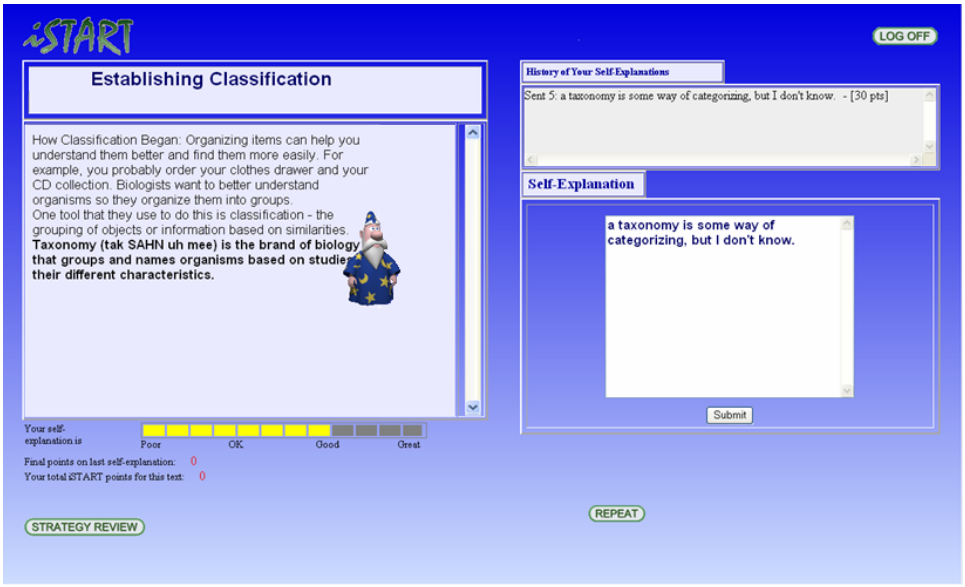
During the introduction module, three animated agents (one teacher and two students) hold a vicarious, classroom-like dialogue. This dialogue presents the concept of SE and the associated iSTART reading strategies (comprehension monitoring, prediction, paraphrasing, elaboration, and bridging). The agents interact with one another to provide descriptions, examples, and counter examples of each comprehension strategy. After each strategy discussion, formative assessments are presented that gauge the student's current level of understanding for that strategy.

After all of the strategies have been introduced and modelled, the system transitions into the demonstration module. The demonstration module utilises two animated agents (one teacher and one student) that apply the SE strategies in the context of an example text. During this scaffolding phase, the user is asked to analyse and identify the various strategies being used by the student agent. The dialogue and feedback between the animated agents foreshadow the interaction that the users will have during the practice module.

The practice module in iSTART affords students the opportunity to apply the iSTART strategies within their own SEs (see Figure 1 for screenshot of the practice environment). This module fades out most direct instruction and uses formative feedback to guide the interaction. Merlin (the teacher agent during demonstration) serves as the SE coach by providing feedback for every student-generated SE and prompting them to use the newly acquired strategies. The main purpose of this module is to provide students with an opportunity to apply the strategies to new, challenging texts and to integrate knowledge from different sources to understand complex content.

During practice, each SE that a student generates is scored by the iSTART assessment algorithm. This assessment helps to inform the feedback provided by the system. The algorithm scoring output is coded as a 0, 1, 2, or 3. An assessment of '0' relates to SEs that are too short or contain mostly irrelevant information. A score of '1' indicates a SE that primarily relates to the target sentence itself (sentence-based). A '2' means that the student incorporated some aspect of the text beyond the target sentence (text-based). If a SE earns a '3', then it has incorporated information at a more global level, and may include outside information or refer to an overall theme (global-based). This algorithm has demonstrated performance comparable to humans, and indicates the general amount of cognitive processing required to generate each SE (Jackson et al., 2010).

Figure 1 Screenshot of coached practice (see online version for colours)



During practice, each SE that a student generates is scored by the iSTART assessment algorithm. This assessment helps to inform the feedback provided by the system. The algorithm scoring output is coded as a 0, 1, 2, or 3. An assessment of '0' relates to SEs that are too short or contain mostly irrelevant information. A score of '1' indicates a SE

that primarily relates to the target sentence itself (sentence-based). A '2' means that the student incorporated some aspect of the text beyond the target sentence (text-based). If a SE earns a '3', then it has incorporated information at a more global level, and may include outside information or refer to an overall theme (global-based). This algorithm has demonstrated performance comparable to humans, and indicates the general amount of cognitive processing required to generate each SE (Jackson et al., 2010).

Within iSTART, there are two types of practice modules. The first practice module is situated within the core context of iSTART (initial two-hour training) and includes two texts. The second practice module is a form of extended interaction, and it operates in the same manner as the original practice module. The extended practice module is designed to provide a long-term learning environment that can span weeks or months. Research on iSTART has shown that the extended practice effectively increases students' performance over time (Jackson et al., 2010). However, one unfortunate side effect of this long-term interaction is that students often become disengaged and uninterested in using the system (Bell and McNamara, 2007; Jackson and McNamara, in press).

3 iSTART-ME

To combat the problem of disengagement over time, the iSTART extended practice module has been situated within a game-based environment called iSTART-ME (Motivationally Enhanced). This game-based environment builds upon the existing iSTART system and was specifically designed to increase persistence and active engagement for students who are more likely to disengage from extended training. The iSTART-ME system and design rationale have been more extensively described in other papers, so only the relevant aspects will be described here (Jackson et al., 2009, 2010).

The main focus of the iSTART-ME project is to implement and assess game-based principles and features that are expected to support effective learning, increase motivation, and sustain engagement throughout a long-term tutorial interaction. Previous research has indicated that increasing self-efficacy, interest, engagement, and self-regulation should positively impact learning (Alexander et al., 1997; Bandura, 2000; Pajares, 1996; Pintrich, 2000; Zimmerman and Schunk, 2001). The iSTART-ME project attempts to manipulate these motivational constructs via game-based features that map onto one of the following five categories: feedback, incentives, task difficulty, control, and environment. These categories are discussed in detail in McNamara et al. (2010).

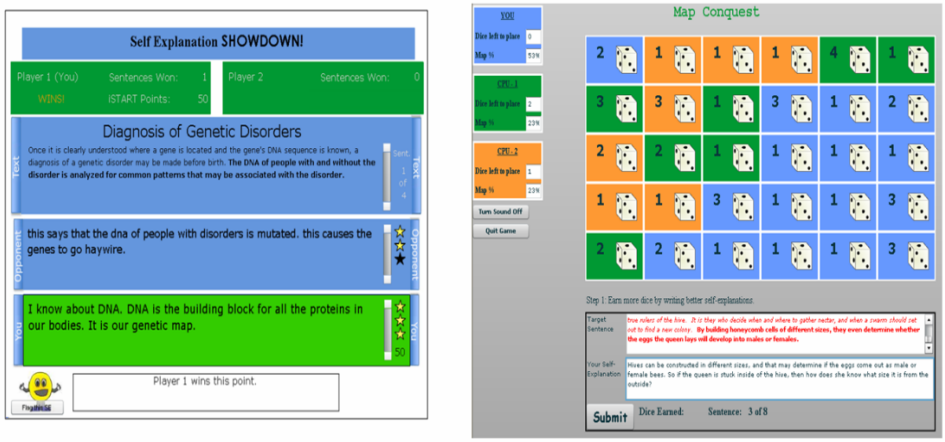
The ITS version of iSTART automatically progresses students from one text to another with no intervening actions. iSTART-ME, however, is controlled through a selection menu (see Figure 2 for screenshot). This selection menu provides students with opportunities to interact with new texts, earn points, advance through levels, purchase rewards, personalise a character, and play educational mini-games (designed to use the same strategies as in practice).

Figure 2 Screenshot of iSTART-ME selection menu (see online version for colours)

Within iSTART-ME, students can earn points as they interact with texts and provide their own SEs in three different generative environments: coached practice, showdown, and map conquest. Coached practice is the same practice environment used within the original version of iSTART (Figure 1). Showdown and map conquest are two methods of generative, game-based practice that use the same iSTART assessment algorithm from regular practice. In showdown [Figure 3(a)], students compete against a computer player to win rounds by writing better SEs. After the learner submits a SE, it is scored and the quality assessment is represented as a number of stars (0–3). The opponent's SE is presented and scored in the same manner. The SE scores are compared and the player with the most stars wins the round. The player who wins the most rounds by the end of a text is declared the overall winner. Map conquest [Figure 3(b)] is the other game-based method of practice where students generate their own SEs. In this game, the quality of a student's SE determines the number of dice that the student earns. Students place these dice on a map and use them to conquer neighbouring opponent territories, which are controlled by two virtual opponents.

The system allows students to freely choose between these three environments for each new text. All students' SEs (regardless of practice environment) are assessed by the iSTART algorithm and points are awarded based on the same scoring rubric. The rubric has been designed to reward consistently good performance. So, students earn more points if they repeatedly provide high-quality SEs on consecutive turns but earn fewer points if they fluctuate between good and poor performance. In addition to providing a form of feedback, the points within iSTART-ME serve three main purposes: advancing through levels, purchasing rewards, and unlocking menu features.

Figure 3 Screenshots of (a) showdown and (b) map conquest (see online version for colours)



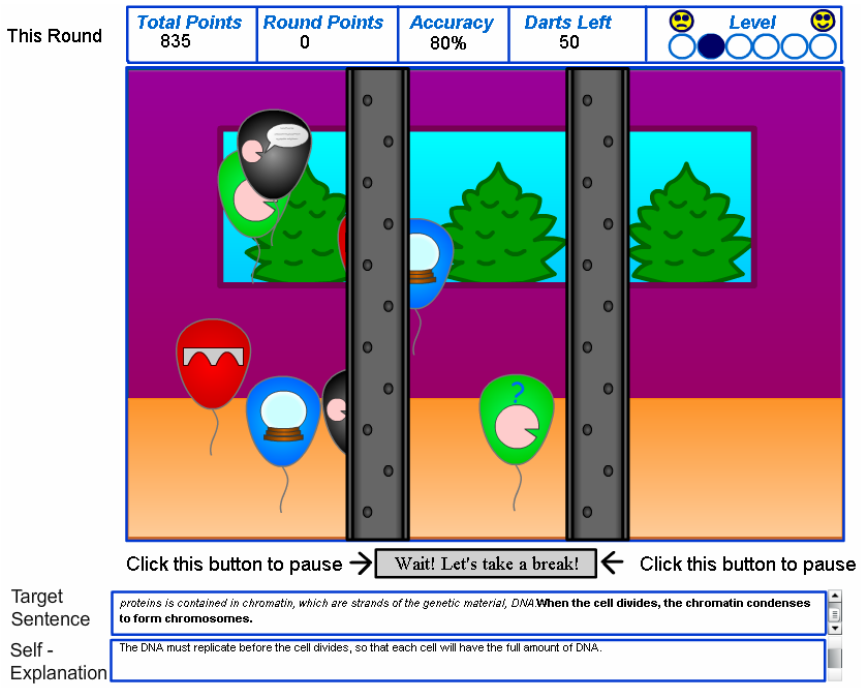
As students accumulate more points, they advance through a series of levels and each new level unlocks one or more new features or games. Each subsequent level requires an increasing number of points; therefore, students must expend slightly more effort to achieve further advancements. The levels are labelled to help increase interest (e.g., 'ultimate bookworm', 'serious strategiser', etc.) and also help to serve as global indicators of progress across texts.

Points can also be used to 'purchase' rewards within the system (bottom box in Figure 2). One of the rewards available to students is the option to change aspects of the learning environment. They can spend some of their iBucks to choose a new tutor agent, change the interface to a new colour scheme, or update the appearance of their personal avatar. These features provide students with a substantial amount of control and personalisation and have been designed as purchasable replacements, rather than always-available options, to help reduce off-task behaviours (such as switching back and forth between agents).

Lastly, a suite of eight educational mini-games has been designed and incorporated within the iSTART-ME extended practice module. Some mini-games require identification of types of strategy use, while others may require students to generate their own SEs. The majority of iSTART-ME mini-games require similar cognitive processes enveloped within different combinations of gaming elements.

In most of the identification mini-games, students are presented with a target sentence and an example SE. For example, in the game balloon bust (Figure 4), students must decide which iSTART strategy was used in the SE and then click on the corresponding balloons. There are three other mini-games that focus on the same task of identifying strategies within example SEs. These other games each incorporate a new interface with a different combination of game elements, including fantasy, competition, and perceptual aspects (as in balloon bust). Though the surface features of these games can differ widely, they have been designed with similar levelling structures and can all be completed within 10–20 minutes. Students are allowed to select any form of practice or mini-game from the selection menu that has been unlocked (provided that they have enough iBucks).

Figure 4 Screenshot of balloon bust (see online version for colours)



4 Current study

Previous research focusing on components of the iSTART-ME system yielded somewhat conflicting patterns of results, depending on the time-scale of the intervention. This previous research indicated that after a short-term interaction (~60 minutes, including brief training), students who used a game-based method of practice performed worse than students using a non-game-based environment (Jackson et al., 2012). However, in a longer-term pilot evaluation with full training (~6 hours across multiple sessions), students performed equally well using either the game-based or the non-game-based practice environments. Therefore, one possible concern with integrating games into learning systems is a potential trade-off between enjoyment and learning (Jackson et al., 2012; Jackson and McNamara, in press). These conflicting results, in conjunction with the previous research on educational games, led us to ask two critical questions related to game-based learning:

1 What is the relative learning benefit from an educational game compared to an ITS?

2 If learning does occur, which students show the most benefits from training?

The current study attempts to address these questions through a multi-session comparison of students' learning gains across three different training conditions: an educational game (iSTART-ME), an ITS (iSTART), and a no-tutoring control. Additionally, we investigate multiple individual difference measures (i.e., domain knowledge, strategy knowledge,

reading ability, and reading habits) to determine which students exhibited the most benefits from these systems.

4.1 Procedure

- *Participants and setting*. High-school students (n = 125) were recruited from the general citywide population in a mid-south urban environment (51% male; 81% African American, 13% Caucasian, 6% other nationalities; average grade completed = 10th grade; average age = 15.8 years). The ten-session experiment was conducted in a university research laboratory and involved three phases: pretest, training, and posttest.

- *Pretest*. During the first session, students completed a pretest that included questions to collect basic demographics (including questions to assess students' reading behaviours), prior reading comprehension, prior strategy knowledge, prior science knowledge, and an assessment of their prior ability to self-explain (relevant details are discussed below).

- *Training*. During each training session, students interacted with their randomly assigned between-subjects condition: a game-based system (iSTART-ME), a traditional ITS (iSTART), or control (delayed training after posttest). Students in the educational game condition interacted with the full game-based selection menu in iSTART-ME across eight separate sessions, lasting a minimum of 1 hour each. Participants in the ITS condition used the original non-game-based version of iSTART for the same amount of time (eight sessions of at least 1 hour each). Students assigned to the control condition completed the pretest and returned a week later for the posttest (their training was delayed until after posttest).

  The initial training within both iSTART systems was identical until the participants transitioned into extended practice. That is, both conditions progressed through the Introduction module, the Demonstration module, and then two regular practice texts within the Coached Practice environment. After these two practice texts, students assigned to the educational game interacted with the full selection menu and chose their own practice environments (Figure 2), while the ITS students practiced by continually transitioning from one text to another within the Coached Practice environment (Figure 1). Both systems allowed students to progress through the tutoring at their own pace; therefore, not all students experienced the same components at the same time. This is a key characteristic of ITSs and games that adapt interactions to user input. Hence, some students naturally receive more or different forms of training and practice than others.

- *Posttest*. All students completed the posttest (session 10), which consisted of assessments similar to those from the pretest. These included measures of SE ability along with questions pertaining to students' attitudes, perceptions, and experiences.

4.2 Measures

- *SE ability*. During the two testing phases, students were presented with a new text (not included within training) and prompted to self-explain specific sentences

(eight SEs during each test). These texts were selected due to their similarity in terms of length (281–329 words), content difficulty (Flesch-Kincaid grade level 8–9), and linguistic features (i.e., similar scores on the five easability component scores within Coh-Metrix (Graesser et al., 2011). Each SE provided by the students was scored using the iSTART assessment algorithm.

- *Reading habits and values*. Questions pertaining to students reading habits included: reading for enjoyment, engaging in extracurricular reading, the number of hours spent reading for science courses, and the number of hours devoted to reading and studying for other classes.

- *Reading comprehension*. Reading comprehension was measured using the 48-item multiple choice Gates-MacGinitie reading comprehension measure (Form K, Level 7/9; MacGinitie and MacGinitie, 1989).

- *Reading strategy knowledge*. Students' reading strategy knowledge was assessed using an adapted and shortened form of the metacognitive reading strategy index (MSI; Schmitt, 1990). The nine-item version of the survey asked students to indicate appropriate behaviours to engage in before, during, and after reading.

- *General science knowledge*. General science knowledge was assessed using a previously validated 20-item, four alternative multiple-choice test addressing knowledge of areas such as biology, scientific methods, mathematics, earth science, and chemistry (McNamara et al., 2006; O'Reilly et al., 2004; O'Reilly and McNamara, 2007).

5 Results

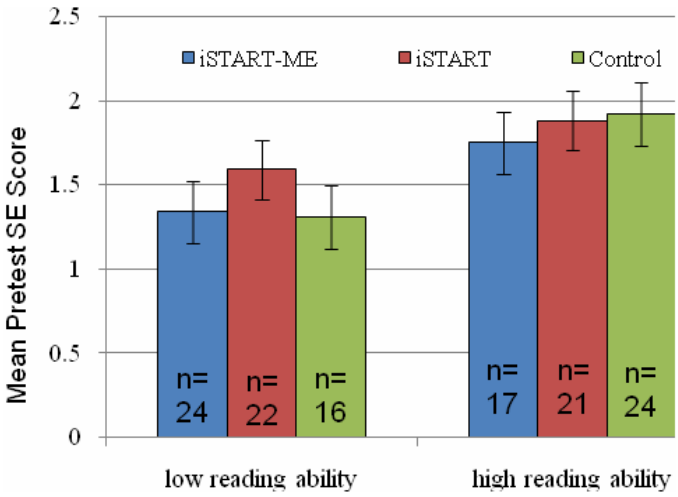
5.1 Learning across conditions and reading abilities

The pretest and posttest SE scores were analysed to assess the potential impacts of the three training conditions and students' prior reading abilities. The following analyses focus on the differences in students' SE quality from pretest to posttest.

A median split using the pretest Gates-MacGinitie reading comprehension test was used to create groups of students with either high or low prior reading ability. Those students in the high reading ability group had a significantly higher proportion of correct answers on the Gates MacGinitie reading comprehension test (n = 63; M = .64) than did the students in the low reading ability group (n = 63, M = .30), $F(1,124) = 260.66$, $p < .001$; $\eta^2 = .678$.

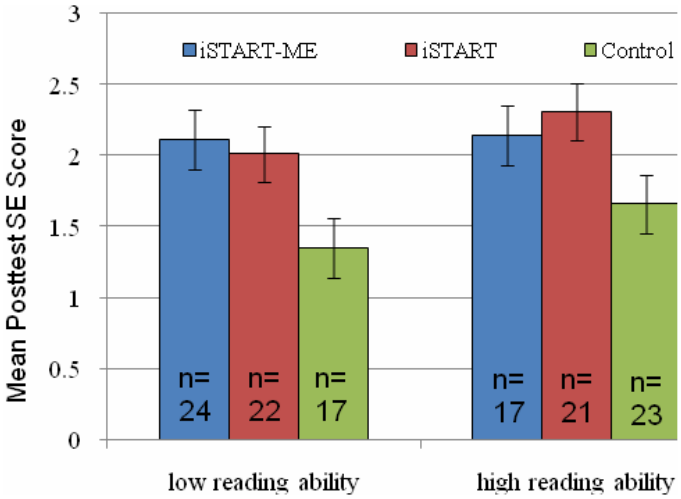
An ANOVA conducted on the pretest SE scores including the between-subjects factors of condition (iSTART, iSTART-ME, control) and reading ability (high, low) indicated that the participants with lower prior reading ability generated significantly lower quality SEs than the high ability students, $F(1,118) = 16.95$, $p < .001$; $\eta^2 = .126$ (see Figure 5 for pretest means). There were no significant differences of pretest SE quality between the three conditions, $F(2,118) = 1.06$, $p = .351$, nor was there a significant interaction between condition and prior reading ability, $F(2,118) = 0.74$, $p = .479$.

Figure 5 Mean pretest SE scores (see online version for colours)



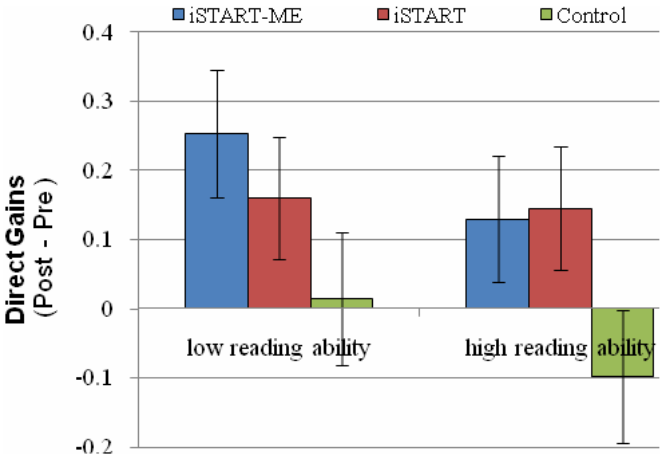
A comparable ANOVA was conducted on the posttest SE quality. As conveyed in Figure 6, there was a significant effect of condition at posttest, $F(2,118) = 13.90$, $p < .001$; $\eta^2 = .191$, revealing the finding that both training conditions (iSTART and iSTART-ME) produced significantly higher quality posttest SEs than the control condition. There was also a marginally significant effect of reading ability indicative that high ability students produced slightly better quality posttest SEs than the participants with low prior reading ability, $F(1,118) = 2.84$, $p = .095$; $\eta^2 = .024$. Notably, the effect of reading ability was marginal and small, and the interaction between condition and reading ability was not significant, $F(2,118) = 0.49$, $p = .617$, indicating that reading strategy training was effective regardless of prior reading ability.

Figure 6 Mean posttest SE scores (see online version for colours)



Two gain scores (direct gain and relative gain) were calculated to further investigate learning differences between conditions. Students' SE performance from pretest and posttest were converted into proportion scores to facilitate comparisons across tests and environments (this allows gains to be measured in terms of percentages rather than raw scores). A *direct gain* score was calculated for each participant by subtracting the pretest SE proportion score from the posttest SE proportion score (see Figure 7 for means).

Figure 7 Direct gains on SE proportion scores (see online version for colours)

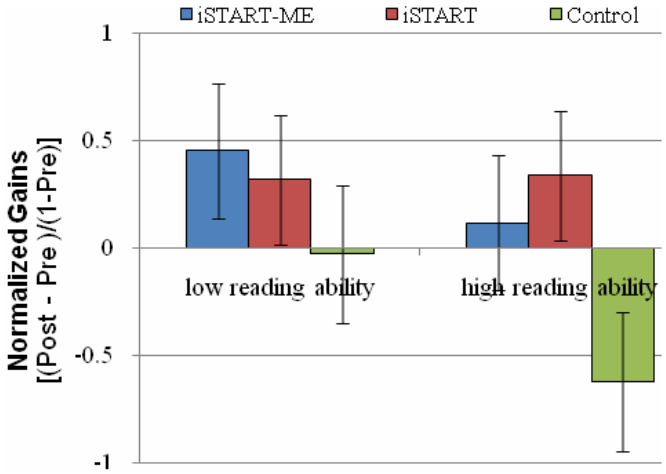


An ANOVA on the direct gains scores yielded a significant main effect for condition, $F(2,116) = 13.39$, $p < .001$; $\eta^2 = .188$, a significant main effect for reading ability, $F(1,116) = 4.74$, $p = .032$; $\eta^2 = .039$, but a non-significant interaction, $F(2,116) = 0.83$, $p = .439$. Using Bonferonni alpha-adjustments for multiple comparisons, we determined that both versions of the iSTART training produced significantly higher direct gains than the control condition ($p < .001$), but were not different from each other ($p = 1.00$). The participants with lower prior reading ability gained slightly more than the high reading ability students ($p = .032$), though they started with lower pretest scores and therefore had more room for improvement. The interaction between condition and reading ability was not significant. However, the iSTART-ME group had a tendency to produce the most gains within the target ability group (i.e., students with low prior ability).

To account for the bias of direct gain scores (i.e., lower people having more room to gain), a *relative gain* score was also calculated. Relative gain scores represent the amount of improvement achieved based on the amount of improvement possible [(Posttest proportion – Pretest proportion) / (1 – Pretest proportion)]. This learning metric reduces the bias towards low ability students and accounts for the difficulty that students may have at improving within the higher levels of performance (see Figure 8 for means). An ANOVA on the relative gains scores revealed a significant main effect for condition, $F(2,116) = 10.11$, $p < .001$; $\eta^2 = .148$, a significant main effect for reading ability, $F(1,116) = 5.45$, $p = .021$; $\eta^2 = .045$, but a non-significant interaction, $F(2,116) = 1.87$, $p = .159$; $\eta^2 = .031$. Using Bonferonni alpha-adjustments for multiple comparisons, we determined that both versions of the iSTART training produced significantly higher relative gains than the control condition ($p < .001$), but were not different from each other ($p = 1.00$). The participants with lower prior reading ability gained slightly more than the

high reading ability students (p = .021), thus replicating the direct gains result using a less biased measure. There was not a significant interaction for relative learning gains between condition and prior reading ability.

Figure 8 Relative gains on SE proportion scores (see online version for colours)



These findings indicate that students in the two training conditions improved their performance significantly beyond the control condition, and students with low prior reading ability improved as much as or more than high ability students. Although these are not surprising effects, they provide a critical baseline for the efficacy of the tutoring systems and establish that game and non-game environments can provide equivalent performance improvements.

5.2 Commitment to reading (or lack thereof)

A series of follow-up analyses were conducted to examine the properties and parameters related to students' SE learning gains across the two tutoring conditions (iSTART and iSTART-ME). Surprisingly, correlation analyses revealed that student learning gains were not related to prior reading strategy knowledge, domain knowledge, or reading ability (see top three rows in Table 1).

In contrast, learning gains were significantly correlated with a series of questions that tap into students' commitment to reading (bottom four rows in Table 1). All of the self-reported reading commitment measures were negatively related to students' learning. Thus, those students who lack a commitment to reading before training are the same students who gain the most from the strategy training provided by the two iSTART systems.

Further analyses were conducted to examine how these reading measures relate to students' SE quality at pretest and posttest. To do so, median splits were used to identify students that reported either high or low levels of reading enjoyment, extracurricular reading, hours devoted to reading for science, and hours devoted to reading for other courses. Figures 9 through 12 display the pretest and posttest SE means for each of these median split groups.

Table 1 Correlations with direct and relative learning gains (n = 84)

Measure	*Direct gains*		*Relative gains*	
	r	*p*	*r*	*p*
Prior strategy knowledge	–.155	(.159)	–.074	(.506)
Prior science knowledge	–.073	(.509)	–.069	(.530)
Prior reading ability	–.153	(.164)	–.098	(.376)
How much do you enjoy reading?	–.226*	(.039)	–.196	(.073)
How many books do you read each year that are not required by your teachers?	–.364*	(.001)	–.318*	(.003)
How many hours *per week* do you devote to reading and studying for your science course?	–.276*	(.011)	–.284*	(.009)
How many hours *per week* do you devote to reading and studying for your other courses (combined) this year	–.254*	(.020)	–.274*	(.012)

Note: *p < .05

A mixed-factor ANOVA was conducted including the between-subjects factors of reading enjoyment (i.e., based on the question, 'How much do you enjoy reading?') and condition (iSTART, iSTART-ME) and the within-subjects factor of testing time. As found in the earlier analyses, SE quality improved from pretest to posttest, $F(1,80) = 47.00$, $p < .001$; $\eta^2 = .370$. However, the effect of condition was not significant, $F(1,80) = 1.37$, $p = .246$ (see Figure 9 for means collapsed across conditions), nor was the interaction between testing time and condition, $F(1,80) = 0.65$, $p = .423$. The same analysis further revealed a non-significant main-effect for reading enjoyment group, $F(1,80) = 0.92$, $p = .340$, but a marginally significant interaction between reading enjoyment and testing time, $F(1,80) = 3.15$, $p = .080$; $\eta^2. = .038$. Two follow-up ANOVA analyses confirmed that participants who tended to enjoy reading less generated lower quality SEs prior to training, $F(1,82) = 3.89$, $p = .052$; $\eta^2 = .045$. However, the two reading enjoyment groups were equivalent at posttest, $F(1,82) = 0.13$, $p = .910$.

Figure 9 How much do you enjoy reading? (see online version for colours)

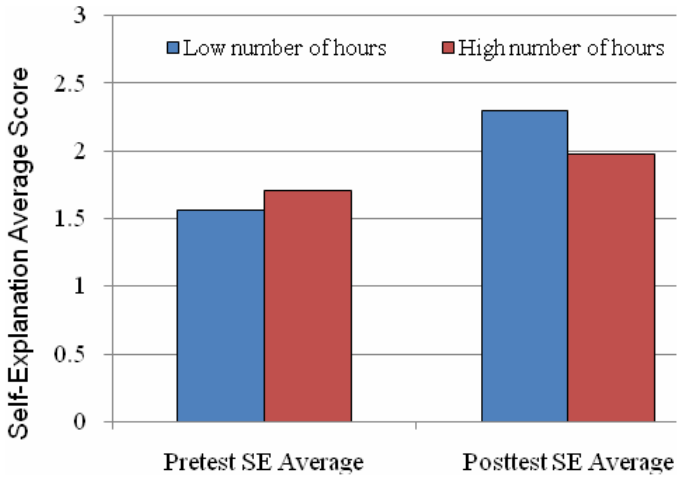
Figure 10 How many books do you read each year that are not required by your teachers? (see online version for colours)

A similar mixed-factor ANOVA (Figure 10) was conducted including the between-subjects factors of extracurricular reading (i.e., based on the question, 'How many books do you read each year that are not required by your teachers?') and condition (iSTART, iSTART-ME) and the within-subjects factor of testing time. There was a significant main effect for testing time, $F(1,80) = 58.46$, $p < .001$; $\eta^2 = .422$, a non-significant main-effect for the extracurricular reading groups, $F(1,80) = 1.13$, $p = .292$, and a significant interaction between testing time and extracurricular reading, $F(1,80) = 8.34$, $p = .005$; $\eta^2 = .094$. There was not a significant main effect of condition, $F(1,80) = 1.65$, $p = .202$, interaction between testing time and condition, $F(1,80) = 1.29$, $p = .260$, or interaction among condition, test time, and extracurricular reading, $F(1,80) = 0.040$, $p = .842$. Follow-up ANOVA analyses indicated that low book readers scored significantly lower than high book readers on pretest SE performance, $F(1,82) = 6.67$, $p = .012$; $\eta^2 = .075$, but that low readers were not different from high readers on posttest SE quality, $F(1,82) = 0.38$, $p = .537$. These results demonstrate that all students benefit from iSTART training (main effect of testing time). The findings also indicate that the training systems provide the most performance benefits for students who are not engaging in extracurricular reading (interaction between testing time and reading group), and help low readers to catch up and match performance of those students who already read more on their own.

A mixed-factor ANOVA (Figure 11) was conducted including the between-subjects factors of science reading time (i.e., based on the question, 'How many hours *per week* do you devote to reading and studying for your science course?') and condition (iSTART, iSTART-ME) and the within-subjects factor of testing time. There was a significant main effect of testing time, $F(1,80) = 52.02$, $p < .001$; $\eta^2 = .394$, a non-significant main-effect for science reading time group, $F(1,80) = 0.55$, $p = .461$, and a significant interaction between testing time and science reading time, $F(1,80) = 10.73$, $p = .002$; $\eta^2 = .118$. There was no significant main effect of condition $F(1,80) = 1.62$, $p = .208$, interaction between condition and testing time, $F(1,80) = 1.36$, $p = .247$, or interaction among

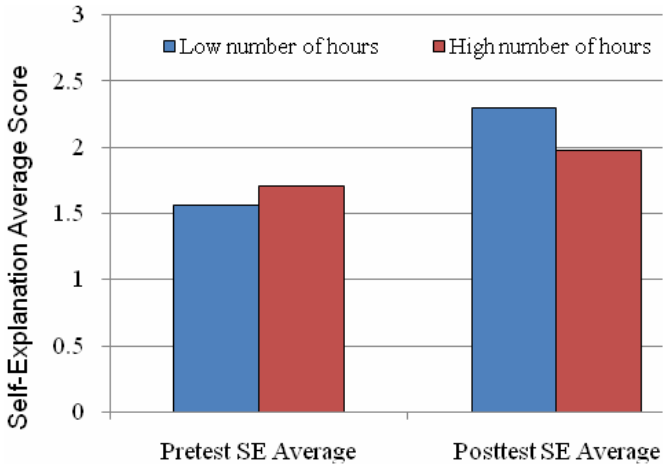
condition, test, and science reading, $F(1,80) = 0.834$, $p = .364$. Follow-up ANOVAs indicated that low science readers scored slightly lower than high science readers on pretest SE performance; however, this difference was not significant, $F(1,82) = 1.35$, $p = .248$. In contrast, an ANOVA revealed that low science readers performed significantly higher than high science readers on posttest SE quality, $F(1,82) = 5.15$, $p = .026$; $\eta^2 = .059$. Again, these results indicate that the iSTART training systems provide the most performance benefits for students who are not reading many hours for science, and help them to catch up and match performance of those students who already read and study more science on their own.

Figure 11 How many hours *per week* do you devote to reading and studying for your science course? (see online version for colours)



A final mixed-factor ANOVA (Figure 12) was conducted including the between-subjects factors of overall reading time (i.e., based on the question, ‘‘How many hours *per week* do you devote to reading and studying for other courses (combined) this year?’) and condition (iSTART, iSTART-ME) and the within-subjects factor of testing time. The analysis yielded a significant main effect for testing time, $F(1,80) = 56.89$, $p < .001$; $\eta^2 = .416$, a non-significant main-effect for overall reading time group, $F(1,80) = 0.00$, $p = .952$, and a significant interaction between overall reading time and testing time, $F(1,80) = 5.48$, $p = .022$; $\eta^2 = .064$. There was no significant main effect of condition, $F(1,80) = 1.64$, $p = .205$, interaction between condition and testing time, $F(1,80) = 1.13$, $p = .291$, or interaction among condition, test, and overall reading time, $F(1,80) = 0.13$, $p = .725$. Follow-up ANOVA analyses confirmed that students who spent less time reading for their courses tended to score lower than high-frequency studiers on pretest SE performance; however, this difference was not significant, $F(1,82) = 1.87$, $p = .175$. Low studiers tended to perform better than high studiers on posttest SE quality, although this is again not a significant difference, $F(1,82) = 1.34$, $p = .250$. These results, though not significant, demonstrate similar patterns and trends to the previous reading questions.

Figure 12 How many hours *per week* do you devote to reading and studying for your other courses (combined) this year? (see online version for colours)



6 Conclusions

The overarching goal of the iSTART-ME project has been to further our understanding of the benefits of designing game-based educational systems. The current efficacy study was a major step in assessing a fully implemented game-based tutoring system built to augment an existing ITS. The primary purpose of this study was to compare the potential learning benefits among three conditions: traditional ITS (iSTART) training, educational game (iSTART-ME) training, and a time-based control. Further, we aimed to identify which students benefited most from strategy training within the two learning environments.

In line with the main research question, results indicated that students learned from both training environments. Specifically, analyses revealed that students' prior reading ability was related to their SE performance at pretest (i.e., higher reading ability yielded higher quality pretest SEs – Figure 5). However, this relation decreased after training (became only marginally significant); the quality of SEs at posttest was instead determined by the randomly assigned training condition, with both iSTART systems outperforming the control condition regardless of initial reading ability (Figure 6). The analyses on direct and relative gains for SE improvement indicated that all students who received training improved above and beyond the control group, as expected (Figures 7 and 8). Further, though the interaction was not significant, the largest improvements in SE quality were found for the low ability readers within the iSTART-ME condition. These results support the main goal of the project, and provide further evidence that games can be effectively integrated within learning environments.

The current long-term evaluation goes well beyond immediate short-term findings to explore the effects of games during prolonged skill acquisition (i.e., using comprehension strategies effectively within SEs). During the extended practice sessions, students in the iSTART-ME condition had access to the full selection menu and therefore could spend more time off-task interacting with various features and mini-games. In contrast, students

in the iSTART condition had continued generative practice and received formative feedback on how to generate higher quality SEs. Despite the potential differences in time-on-task allocation, the students using the game-based system gained equivalently to students using the traditional ITS. This finding supports the long-term learning trend from previous work (Jackson et al., 2012) and creates a promising foundation that can support subsequent work and contribute to the research on game-based learning.

In addition to the learning gain comparisons between conditions, follow-up analyses focused on identifying which students benefitted most from strategy training. Correlations indicated that many of the individual difference measures (i.e., prior strategy knowledge, domain knowledge, and reading ability) were not significantly related to students' learning gains (Table 1). Instead, the analyses indicated that learning gains were negatively related to students' commitment to reading. Thus, students who were less committed to reading prior to the study produced the largest learning gains. To further explore this relation, a series of statistical comparisons (Figures 9 to 12) found that students who began the study with a low commitment to reading benefited the most from training, as evidenced by their posttest performance catching up to or exceeding the performance level of highly committed readers. This finding held true for four separate questions, which all related to students' reading habits and values: reading enjoyment, extracurricular reading, hours spent studying for a science course, and hours spent studying for all other courses. These same analyses found no significant differences between the two training systems, indicating that game-based training can produce learning gains equivalent to a traditional ITS. These findings add to previous research using the ITS version of iSTART, which found that during extended strategy practice, low prior reading ability students were able to catch-up and match performance of students with a high prior reading ability (Jackson et al., 2010).

Overall, these results support the primary goal of the iSTART systems; namely, they suggest that comprehension strategy instruction and practice can help students overcome initial skill deficits and compensate for prior individual differences. These findings are particularly relevant for systems and educators complying with the content literacy requirements established within the Common Core State Standards. Based on the results, iSTART training should provide significant benefits for struggling readers and improve their ability to understand complex content.

The development of iSTART-ME allows us to examine the effectiveness of an educational game, as well as to more systematically evaluate the effects of game components in the context of an ITS. The current study compared two separate systems and their relation to student learning. However, iSTART-ME has been designed with distinct and separable features so that multiple system configurations can be easily implemented and tested across a variety of experiments. Future smaller-scale experiments have been designed to leverage this modular design in order to examine the interactions among students' individual differences and combinations of system components (e.g., presence or absence of avatars, trophies, mini-games), as well as changes in user performance, enjoyment, attitudes, engagement, and persistence across time.

Our research using iSTART and iSTART-ME is intended to improve our understanding of techniques that foster content-text comprehension as well as enhance the design of ITSs and game-based learning environments. Our overarching goal is to better understand the complex interplay between motivation and learning, and contribute to research on how to most effectively combine ITS and game-based principles.

Ultimately, we expect hybrid game-based tutoring environments to dramatically impact the effectiveness of computer-based training as well as further our understanding of the complex motivational aspects of educational systems and their interplay with learning.

Acknowledgements

This research was supported in part by the Institute for Educational Sciences (IES R305G020018-02; R305G040046; R305A080589) and National Science Foundation (NSF REC0241144; IIS-0735682). Any opinions, findings, or recommendations in this material are those of the authors and do not necessarily reflect the views of IES or NSF.

References

Alexander, P., Murphy, P., Woods, B., Duhon, K. and Parker, D. (1997) 'College instruction and concomitant changes in students' knowledge, interest, and strategy use: a study of domain learning', *Contemporary Educational Psychology*, Vol. 22, No. 2, pp.125–146.

Anderson, J., Corbett, A., Koedinger, K. and Pelletier, R. (1995) 'Cognitive tutors: lessons learned', *The Journal of Learning Sciences*, Vol. 4, No. 2, pp.167–207.

Bandura, A. (2000) 'Self-efficacy: The foundation of agency', in Perig, W. and Grob, A (Eds.): *Control of Human Behavior, Mental Processes, and Consciousness: Essays in Honor of the 60th Birthday of August Flammer*, pp.17–33, Erlbaum, Mahwah, NJ.

Bell, C. and McNamara, D. (2007) 'Integrating iSTART into a high school curriculum', *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, pp.809–814, Cognitive Science Society, Austin, TX.

Boekaerts, M., Pintrich, P. and Zeidner, M. (Eds.) (2000) *Handbook of Self-regulation*, Academic Press, San Diego, CA.

Clark, D., Nelson, B., Sengupta, P. and D'Angelo, C. (2009) *Rethinking Science Learning through Digital Games and Simulations: Genres, Examples, and Evidence*, National Research Council, Washington DC.

Corbett, A. and Anderson, J. (1990) 'The effect of feedback control on learning to program with the LISP tutor', in *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp.796–803, Cambridge, MA.

Craig, S., D'Mello, S., Gholson, B., Witherspoon, A., Sullins, J. and Graesser, A. (2004) 'Emotions during learning: the first steps toward an affect sensitive intelligent tutoring system', in *Proceedings of E-learn 2004: World Conference on E-learning in Corporate, Government, Healthcare, and Higher Education*, pp.284–288, AACE, Chesapeake, VA.

D'Mello, S. and Graesser, A. (2006) 'Affect detection from human-computer dialogue with an intelligent tutoring system', in Gratch, J. et al. (Eds.): *IVA 2006*, pp.54–67, LNAI 4133, Springer-Verlag, Berlin, Heiderlberg.

D'Mello, S., Taylor, R. and Graesser, A. (2007) 'Monitoring affective trajectories during complex learning', in *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, pp.203–208, Cognitive Science Society, Auxtin, TX.

Foltz, P., Gilliam, S. and Kendall, S. (2000) 'Supporting content-based feedback in online writing evaluation with LSA', *Interactive Learning Environments*, Vol. 8, No. 2, pp.111–129.

Graesser, A., McNamara, D., and Kulikowich, J. (2011) 'CohMetrix: providing multilevel analyses of text characteristics', *Educational Researcher*, Vol. 40, No. 5, pp.223–234.

Gredler, M. (2004) 'Games and simulations and their relationships to learning', in Jonassen, D. (Ed.): *Handbook of Research on Educational Communications and Technology*, pp.571–582, Lawrence Erlbaum Associates, Mahwah, NJ.

Jackson, G. and McNamara, D. (in press) 'Motivation and performance in a game-based intelligent tutoring system', *Journal of Educational Psychology*, No. 4.

Jackson, G., Boonthum, C. and McNamara, D. (2009) 'iSTART-ME: situating extended learning within a game-based environment', in *Proceedings of the Workshop on Intelligent Educational Games at the 14th Annual Conference on Artificial Intelligence in Education*, pp.59–68, AIED, Brighton, UK.

Jackson, G., Boonthum, C. and McNamara, D. (2010) 'The efficacy of iSTART extended practice: Low ability students catch up', in *Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, pp.349–351, Springer, Berlin.

Jackson, G., Dempsey K. and McNamara, D. (2010) 'The evolution of an automated reading strategy tutor: from classroom to a game-enhanced automated system', in Khine, M. and Saleh, I. (Eds.): *New Science of Learning: Cognition, Computers and Collaboration in Education*, pp.283–306, Springer, New York, NY.

Jackson, G., Dempsey, K. and McNamara, D.S. (2012) 'Game-based practice in a reading strategy tutoring system: showdown in iSTART-ME', in Reinders, H. (Ed.): *Computer Games, Multilingual Matters*, pp.115–138, Bristol, UK.

Jackson, G., Guess, R. and McNamara, D. (2010) 'Assessing cognitively complex strategy use in an untrained domain', *Topics in Cognitive Science*, Vol. 2, No. 1, pp.127–137.

MacGinitie, W.H. and MacGinitie, R.K. (1989) *Gates-MacGinitie Reading Tests*, 3rd ed., Riverside, Itasca, IL.

Magliano, J., Todaro, S., Millis, K., Wiemer-Hastings, K., Kim, H. and McNamara, D. (2005) 'Changes in reading strategies as a function of reading training: a comparison of live and computerized training', *Journal of Educational Computing Research*, Vol. 32, No. 2, pp.185–208.

McNamara, D. (2004) 'SERT: self-explanation reading training', *Discourse Processes*, Vol. 38, No. 1, pp.1–30.

McNamara, D., Jackson, G. and Graesser, A. (2010) 'Intelligent tutoring and games (ITaG)', in Baek, Y. (Ed.): *Gaming for Classroom-based Learning: Digital Role-playing as a Motivator of Study*, pp.44–65, IGI Global, Hershey, PA.

McNamara, D., O'Reilly, T., Best, R. and Ozuru, Y. (2006) 'Improving adolescent students' reading comprehension with iSTART', *Journal of Educational Computing Research*, Vol. 34, No. 2, pp.147–171.

Newell, A. and Rosenbloom, P. (1981) 'Mechanisms of skill acquisition and the law of practice', in Anderson, J. (Ed.): *Cognitive Skills and Their Acquisition*, pp.1–55, Hillsdale, NJ.

O'Neil, H., Wainess, R. and Baker, E. (2005) 'Classification of learning outcomes: evidence from the computer games literature', *Curriculum Journal*, Vol. 16, No. 4, pp.455–474.

O'Reilly, T. and McNamara, D. (2007) 'The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional 'high-stakes' measures of high school students' science achievement', *American Educational Research Journal*, Vol. 44, No. 1, pp.161–196.

O'Reilly, T., Sinclair, G. and McNamara, D. (2004) 'iSTART: a web-based reading strategy intervention that improves students' science comprehension' in *Proceedings of the IADIS International Conference Cognition and Exploratory Learning in Digital Age: CELDA 2004*, pp.173–180, IADIS Press, Lisbon, Portugal.

O'Reilly, T., Taylor, R. and McNamara, D. (2006) 'Classroom based reading strategy training: self-explanation vs. reading control', in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp.1887–1892, Erlbaum, Mahwah, NJ.

Orbach, E. (1979) 'Simulation games and motivation for learning: a theoretical framework', *Simulation and Games*, Vol. 10, No. 1, pp.3–40.

Pajares, F. (1996) 'Self-efficacy beliefs in academic settings', *Review of Educational Research*, Vol. 66, No. 4, pp.543–578.

Pintrich, P. (2000) 'Multiple goals, multiple pathways: the role of goal orientation in learning and achievement', *Journal of Educational Psychology*, Vol. 92, No. 3, pp.544–555.

Ricci, K., Salas, E. and Cannon-Bowers, J. (1996) 'Do computer-based games facilitate knowledge acquisition and retention?', *Military Psychology*, Vol. 8, No. 4, pp.295–307.

Rodrigo, M., Baker, R., Lagud, M., Lim, S., Macapanpan, A., Pascua, S., Santillano, J., Sevilla, L., Sugay, J., Tep, S. and Viehland, N. (2007) 'Affect and usage choices in simulation problem solving environments', in *Proceedings of Artificial Intelligence in Education*, pp.145–152, AIED, Marina del Rey, CA.

Rowe, J., Shores, L., Mott, B. and Lester, J. (2011) 'Integrating learning, problem solving, and engagement in narrative-centered learning environments', *International Journal of Artificial Intelligence in Education*, Vol. 21, Nos. 1–2, pp.115–133.

Schmitt, M. (1990) 'A questionnaire to measure children's awareness of strategic reading processes', *Reading Teacher*, Vol. 43, No. 7, pp.454–461.

Shank, R. and Neaman, A. (2001) 'Motivation and failure in educational systems design', in Forbus, K. and Feltovich, P. (Eds.): *Smart Machines in Education*, AAAI Press/ MIT Press, Cambridge, MA.

Swaak, J. and de Jong, T. (2001) 'Discovery simulations and the assessment of intuitive knowledge', *Journal of Computer Assisted Learning*, Vol. 17, No. 3, pp.284–295.

VanLehn, K. (2011) 'The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems', *Educational Psychologist*, Vol. 46, No. 4, pp.197–221.

Vogel, J., Vogel, D., Cannon-Bowers, J., Bowers, C., Muse, K. and Wright, M. (2006) 'Computer gaming and interactive simulations for learning: a meta-analysis', *Journal of Educational Computing Research*, Vol. 34, No. 3, pp.229–243.

Wilson, K., Bedwell, W., Lazzara, E., Salas, E., Burke, S., Estock, J., Orvis, K. and Conkey, C. (2009) 'Relationships between game attributes and learning outcomes: review and research proposals', *Simulation and Gaming*, Vol. 40, No. 2, pp.217–266.

Zimmerman, B. and Schunk, D. (2001) 'Reflections on theories of self- regulated learning and academic achievement' in Zimmerman, B. and Schunk, D. (Eds.): *Self-regulated Learning and Academic Achievement: Theoretical Perspectives*, pp.289–307, Erlbaum, Mahwah, NJ.