Motivation and Performance in a Game-Based Intelligent Tutoring System

G. Tanner Jackson and Danielle S. McNamara

Online First Publication, September 9, 2013. doi: 10.1037/a0032580

CITATION
Motivation and Performance in a Game-Based Intelligent Tutoring System

G. Tanner Jackson and Danielle S. McNamara
Arizona State University

One strength of educational games stems from their potential to increase students’ motivation and engagement during educational tasks. However, game features may also detract from principle learning goals and interfere with students’ ability to master the target material. To assess the potential impact of game-based learning environments, in this study we examined motivation and learning for 84 high-school students across eight 1-hr sessions comparing 2 versions of a reading strategy tutoring system, an intelligent tutoring system (iSTART) and its game-based version (iSTART–ME). The results demonstrate equivalent target task performance (i.e., learning) across environments at pretest, posttest, and retention, but significantly higher levels of enjoyment and motivation for the game-based system. Analyses of performance across sessions reveal an initial decrease in performance followed by improvement within the game-based training condition. These results suggest possible constraints and benefits of game-based training, including time-scale effects. The findings from this study offer a potential explanation for some of the mixed findings within the literature and support the integration of game-based features within intelligent tutoring environments that require long-term interactions for students to develop skill mastery.

Keywords: educational games, intelligent tutoring, motivation and performance

Intelligent tutoring systems (ITSs) are automated tutoring environments that adapt to users based on various well-established cognitive principles and algorithms (Anderson, 1982). This approach has been highly successful for the last several decades as evidenced by significant learning gains in studies covering a wide range of domains (e.g., Cohen, Kulik, & Kulik, 1982; Graesser, McNamara, & VanLehn, 2005; Merrill, Reiser, Ranney, & Trafton, 1992). However, one potential weakness of long-term ITSs is that while novel to students at first, they can become repetitive over time. This facet is a particular problem when the targeted skill or knowledge requires extended practice to reach sufficient mastery or depth of understanding.

An increasing number of long-term tutoring systems focus on prolonged skill acquisition across multiple interactions, and several of these have been integrated and evaluated within ecological settings (Jackson, Boothum, & McNamara, 2010; Johnson & Valente, 2008; Koedinger & Corbett, 2006; Meyer & Wijckumar, 2011). Due to the extended time span of these interactions, students can sometimes become disengaged and bored while using some systems (e.g., Arroyo et al., 2007; Baker, D’Mello, Rodrigo, & Graesser, 2010; Bell & McNamara, 2007). When the learning process is expected to require multiple days, weeks, or months, designing the environment such that it induces the learner to persist should be paramount among the design objectives. If students do not remain engaged and persist within a training environment, attaining a long-term learning objective is nearly impossible. Furthermore, for those students who continue to interact despite lack of interest, boredom may trigger a vicious cycle that prevents them from actively re-engaging in constructive learning processes (Baker, Corbett, & Koedinger, 2004; D’Mello, Taylor, & Graesser, 2007).

Educational systems that require a longer training commitment may benefit from design features that enhance student engagement after any novelty effects have dissipated. Nonetheless, there is more to learning than interest and engagement. Sacrificing essential pedagogical aspects of an educational environment to increase interest is not likely to be successful. As these constraints have become more evident, system designers have begun to carefully incorporate educational games and game-based elements to help capture students’ interest and promote active participation within learning environments (McNamara, Jackson, & Graesser, 2010).

Game-Based Learning

It is intuitively clear that games are a potentially strong motivating factor for students (Gee, 2003; Steinkuehler, 2006). A natural, intrinsic interest in the domain content of the system is, of course, the preferred method of obtaining involvement, but unfortunately not all learners share interests. While the content itself plays an important role for determining interest, perhaps the framing of this content (e.g., incorporating it within a game) is even more crucial. Thus, a game itself can be used as a catalyst to promote motivation and sustain the interest of students.

The increased focus on games in education may also be partially due to the alignment between aspects of game design and the goals
of educational environments. This is not simply a grafting of two successful but incompatible technologies; research suggests that these technologies have a common theoretical foundation and that the sum is greater than the parts (e.g., Laird & van Lent, 2000; Van Eck, 2006). Specifically, an essential overarching benefit to games is that they, similar to tutoring systems, provide the opportunity for adaptive, individualized interactions. The notion behind these highly interactive educational games is to involve learners, giving them opportunities to perform, experience outcomes, and reflect on the targeted tasks such that these actions are integrated within a meaningful context (Barab et al., 2010).

Games often improve engagement and motivation by employing features similar to those found within successful tutoring systems. For example, one of the many motivating factors of games is the individual and personalized nature of the interactions that adapt to the skills and actions of the player (Gee, 2005; Malone & Lepper, 1987; Rieber, 1996). To accomplish this goal, an educational game must be able to identify the ability level of the learner and adjust itself accordingly (Conati, 2002; Rieber, 1996; Shute & Towlle, 2003). As such, the game may require demonstration of more advanced skills or knowledge from a learner progressing successfully through the game or lessen the requirements for a learner progressing poorly. Additionally, the rapid feedback within educational games can help learners to better regulate their progress and activities. Indeed, the role of feedback in any learning environment can lend a stronghold on engagement (Anderson, Corbett, Koedinger, & Pelletier, 1995; Corbett & Anderson, 1990; Foltz, Gilliam, & Kendall, 2000). By leveraging these features to increase engagement and motivation, these games are highly compatible with the sophisticated pedagogy implemented within most ITSs.

Another aspect of games that maps onto pedagogical goals is the notion of challenge (i.e., task difficulty; Gredler, 2004; Rieber, 1996). Games that are easily won require little effort from learners. On the other hand, games that are too difficult can result in lowered interest because learners are unable to accomplish goals. Vygotsky (1978) posited that learning is most effective when the material is slightly more advanced than the learner. With respect to game challenge, the same hypothesis could apply. A game that is slightly more challenging than the learner’s skill and knowledge may sustain interest and motivation by providing accomplishment while maintaining effort (Gee, 2003). Indeed, self-efficacy and interest in games have been found to be highly correlated (Zimmerman & Kitsantas, 1997). Ratings of higher self-efficacy during game play coincide with higher preferences for one game over others. Thus, accomplishment by the players over consistent challenges should raise their self-efficacy, overall enjoyment, and motivation to perform the task.

Motivation and Mastery in Educational Games

Ample research shows that learning (and mastery) is more than just a cognitive process (du Boulay, 2011); learning is as much a motivational and affective task as it is a demonstration of mental ability. Research also suggests that there is an indirect link between motivation and learning (Garris, Ahlers, & Driskell, 2002); namely, motivation influences the learning processes in which students engage. And, these processes subsequently affect learning outcomes.

Motivation is a multidimensional construct that subsumes a number of component factors, such as interest, enjoyment, expectations, and values. For the current work, motivation generally refers to students’ desire to perform a task and willingness to expend effort on that activity (Garris et al., 2002; Pintrich & Schrauben, 1992; Wolters, 1998). This broad conceptualization of motivation encompasses previous research examining both intrinsic and extrinsic factors related to interest, engagement, enjoyment, and self-efficacy. This prior work has demonstrated that enhancing these aspects of motivation positively impacts learning (Alexander, Murphy, Woods, Duhan, & Parker, 1997; Bandura, 2000; Pajares, 1996; Pintrich, 2000; Young et al., 2012; Zimmerman & Schunk, 2001). Other research has demonstrated that various mechanisms common to games, such as feedback, incentives, task difficulty, and control, can have a significant impact on these motivational constructs and, hence, may ultimately affect learning (Conati, 2002; Corbett & Anderson, 2001; Cordova & Lepper, 1996; Graesser, Chipman, Leeming, & Biedenbach, 2009; Malone & Lepper, 1987; Moreno & Mayer, 2005; Shute, 2008).

Many games leverage these mechanisms and other features as part of a core game design. No individual feature is required within a game, and some game elements may even be unnecessary, ineffective, or distracting in the short term, but they may also have the potential to increase interest, enjoyment, and engagement in the long term. Previous research has also suggested that the affective benefits from games may increase as the number of incorporated game-based features increases (Cordova & Lepper, 1996; Papastergiou, 2009). Therefore, some researchers have assumed that combining several game features together will provide students with a more enjoyable interaction (Asgari & Kaufman, 2004; McNamara et al., 2010).

Unfortunately, despite the increase in research related to educational games, there remains a dearth of research in which the effectiveness of these new gaming environments have been directly compared with their natural counterpart, traditional intelligent tutoring environments (O’Neil & Fisher, 2004; O’Neil, Wainess, & Baker, 2005). Two recent studies have been conducted in which researchers have directly investigated the effectiveness and benefits of educational game components compared with an ITS. The first study by Jackson, Dempsey, and McNamara (2012) was a 90-min experiment to compare the short-term practice effects of a traditional ITS environment with a game-based counterpart. They found that participants who had interacted with game-based practice rated it as significantly more engaging than students within the traditional ITS. By contrast, students who interacted with the traditional ITS outperformed students who practiced using the game environment.

A second smaller study was conducted over a longer time span (six separate sessions) to investigate a combined system that allowed users to continually choose between practicing with an ITS or a game-based system (Jackson, Dempsey, Graesser, & McNamara, 2011). Participants in this study completed a 2-hr introductory training session before entering the practice environment where they could choose between systems (for the remaining 4–5 hr across sessions). Focusing on the results comparing the same two systems from Jackson et al. (2012), there were no advantages for the traditional ITS in this longer term study in terms of performance (comparing within-subjects). The students performed equally well within both systems. In addition, although
there were trends showing improved enjoyment for the game-based system over the ITS, the difference was not statistically significant. The findings from these studies helped to motivate the current study, provide support for differing hypotheses (discussed in more detail later), and suggest that the current study is needed to further explore the complex interplay between games and learning (also see Harris, 2008).

The current work aims to more directly address these issues in game-based learning by comparing the outcomes from two similar long-term skill acquisition systems: a traditional ITS (iSTART) and an educational game (iSTART–ME).

**iSTART and iSTART–ME**

The Interactive Strategy Training for Active Reading and Thinking-Motivationally Enhanced (iSTART–ME) tutor is a newly developed game-based learning environment built on top of an existing tutoring system (iSTART). iSTART provides young adolescents to college-age students with comprehension strategy training to better understand and learn from challenging science texts (McNamara, Levinstein, & Boonthum, 2004; McNamara, O’Reilly, Best, & Ozuru, 2006). In iSTART, pedagogical agents instruct trainees in the use of self-explanation and other active reading comprehension strategies to explain the meaning of science text while they read. The training was motivated by empirical findings showing that students who self-explain text are more successful at solving problems, more likely to generate inferences, construct more coherent mental models, and develop a deeper understanding of the concepts covered in text (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994; McNamara, 2004).

**iSTART Modules**

Strategy instruction occurs in three stages with each stage requiring increased interaction on the part of the learner. During the Introduction Module of iSTART, a trio of animated characters introduces students to the concept of self-explanation and associated reading strategies by providing information, posing questions, and discussing examples. In the second phase, called the Demonstration Module, two agents demonstrate the use of self-explanation using a science text, and the trainee identifies the strategies being used by the agents. During this module, the teacher character (Merlin) asks the trainee to indicate which strategies the student agent (Genie) employed in producing his self-explanation. Finally, Merlin gives Genie feedback on the quality of his self-explanation.

In the third phase (Practice), Merlin coaches and provides feedback while the trainee practices self-explanation using the repertoire of reading strategies. The goal is to help the trainee acquire the skills necessary to integrate prior text and prior knowledge with the current sentence content. For each sentence, Merlin reads the sentence, asks the trainee to explain it by typing a self-explanation, and provides feedback on the quality of the explanation.

The iSTART assessment algorithm drives the feedback provided by Merlin. The algorithm output is coded as a 0, 1, 2, or 3. An assessment of 0 indicates that the self-explanation was either too short or contained mostly irrelevant information. An iSTART score of 1 is associated with a self-explanation that primarily relates only to the target sentence itself (sentence-based). A 2 means that the student’s self-explanation incorporated some aspect of the text beyond the target sentence (text-based). If a self-explanation earns a 3, then it is interpreted to have incorporated information at a global level and may include outside information.
or refer to an overall theme across the whole text (global-based). This algorithm has demonstrated performance comparable to that of humans and provides a general indication of the cognitive processing required to generate a self-explanation (Jackson, Guess, & McNamara, 2010).

Within iSTART, there are two types of practice modules. The first practice module is situated within the core context of iSTART (initial 2-hr training) and includes two texts. The second practice module is a form of extended practice, which operates in the same manner as the regular practice module. This extended practice phase (called Coached Practice—see Figure 1 for a screenshot) is designed to provide a long-term learning environment that can span weeks or months. Research has shown that this extended practice increases students' performance over time (Jackson, Boonthum, 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 & McNamara, 2007).

**iSTART–ME**

Previous research with iSTART pointed to the need for students to persist within the system across several days of training. Therefore, changes were implemented within the system to combat the problem of disengagement over time. The extended practice module of iSTART was redesigned and situated within a game-based environment called iSTART–ME (Motivationally Enhanced). This game-based environment was built directly on top of the existing iSTART system. The main goal of the iSTART–ME project was to implement several of the game-based principles and mechanisms that were expected to support effective learning, increase motivation, and sustain engagement throughout a long-term interaction with an established ITS. The project attempted to implement and potentially manipulate these motivational constructs via game-based features that map onto one of five interaction mechanisms: feedback, incentives, task difficulty, control, and environment (see McNamara et al., 2010, for more details on the mechanisms).

The original ITS version of iSTART with Coached Practice automatically progresses students from one text to another with no intervening tasks. The new version of iSTART–ME is situated within a cohesive meta-game and point-based economy that the user can control through a selection menu (see Figure 2 for screenshot). This new selection menu provides students with opportunities to interact with new texts (control/task difficulty), earn points and trophies (feedback/incentives), advance through levels (feedback/incentives), unlock new features (control/incentives/environment), personalize a character (control/incentives/environment), and play educational mini-games (control/incentives/task difficulty).

Within iSTART–ME, students earn points as they interact with texts and provide their own self-explanations. Each time a student submits a self-explanation, it is assessed by the iSTART algorithm and points are awarded based on a scoring rubric. The rubric has been designed to reward consistently good performance. So students earn more points if they repeatedly provide good self-explanations but earn fewer points if they fluctuate between good and poor performance. These points help go beyond the qualitative responses from the animated agents to provide an additional, quantifiable form of feedback as students learn and practice the self-explanation strategies. For example, students can easily understand that a score of 30 is better than a score of 10, but it is more difficult to gauge the relative difference between, “All right, let’s keep going” and “You’re starting to get the hang of this.” In addition to serving as a form of feedback, earning points within iSTART–ME serves two main incentive purposes: advancing through levels and purchasing rewards.

As students accumulate more points, they advance through a series of levels. Each subsequent level requires an increasing number of points. Therefore, students must expend slightly more time or effort for further advancement (i.e., increasing task difficulty to reach a new level). Whenever students advance up a level, a new subset of features is automatically unlocked and becomes available within the interface (thus acting as an incentive and providing additional control). Each of the iSTART levels are labeled (e.g., ultimate bookworm, serious strategist) to help provide incentive, increase interest, and serve as global indicators of progress across texts.

Points can also be used to control the environment by “purchasing” incentives within the system. One of the options available as a reward allows students to change aspects of the learning

---

**Figure 2.** Screenshot of iSTART–ME selection menu.
environment. They can spend some of their points to choose a new tutor agent, change the interface to a new color scheme, or update the appearance of their personalizable avatar. These features provide students with a substantial amount of control and personalization over their environment and have been designed as purchasable replacements, rather than continuously available options, to help reduce off-task behaviors (such as switching back and forth between agents solely to see what they all look like).

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

Showdown and Map Conquest are two methods of generative game-based practice that use the same iSTART assessment algorithm from regular practice. In Showdown (see Figure 3 for a screenshot), students compete against a computer player to win rounds by producing better self-explanations. After the learner submits a self-explanation, it is scored, the quality assessment is represented as a number of stars (0–3), and an opponent self-explanation is also presented and scored. The difficulty of the opponent self-explanation has been manipulated within previous experiments (Dempsey, 2011); however, for normal gameplay, the opponent example is chosen at random to provide a range of student modeling (i.e., good and bad examples). The self-explanation scores are compared, and the player with the most stars wins the round. The player with the most rounds at the end of the text is declared the winner. The combination of features for Showdown incorporates aspects of feedback (points, stars, rounds won), incentives (points, stars), control (production of self-explanation), and task difficulty (opponent, text content).

Map Conquest is the other game-based method of practice where students generate their own self-explanations. Within Map Conquest, the quality of a student’s self-explanation determines the number of dice that student earns (i.e., performance at the target task determines the resources available during a subsequent game task). Students place these dice on a map and use them to conquer neighboring opponent territories, which are controlled by two virtual opponents. The surface components are somewhat different from those in Showdown but were similarly designed to provide the user with feedback (points, dice), incentives (dice, map puzzle), control (map puzzle, production of self-explanation), and task difficulty (opponents, text content).

In most of the identification mini-games—for example, Balloon Bust (Figure 4)—students are presented with a target sentence and an example self-explanation. The student must decide which iSTART strategy was used in the self-explanation and then click on the corresponding balloons. There are also three other mini-games that focus on the same task of identifying strategies within example self-explanations. These other games each incorporate a new interface with a different combination of game elements, which might include fantasy, competition, and perceptual aspects (as in Balloon Bust). Though the surface features of these games can differ widely, they have been designed with very similar underlying mechanisms and can all be completed within 10–20 min. 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 points). After completion of a task, students are directed back to the main iSTART–ME selection screen.
Current Study

The current study was a multisession experiment in which the effectiveness of a game-based tutoring system (iSTART–ME) was compared with its ITS counterpart (iSTART–Regular). One possible concern with integrating games into learning systems is that they have the potential to detract from the immediate pedagogical goals and reduce learning improvements in the short term (Jackson et al., 2012; Mayer & Moreno, 2003; Paas, Renkl, & Sweller, 2003). However, across long-term training, the engagement fostered by the game environment may compensate for any distracting elements, thus allowing students to catch up in performance (Jackson, Dempsey, Graesser, & McNamara, 2011). Hence, this study was conducted to thoroughly explore the potential long-term benefits of game-based training, how it compares with training from a traditional ITS, and how various effects of motivation and learning may unfold over time.

Hypotheses

The Jackson et al. (2012) study indicated that students who received game-based training during early stages of skill acquisition exhibited decreased performance at the target task (compared with students in a traditional ITS). In contrast, the Jackson, Dempsey, et al. (2011) study showed that when students completed initial strategy training within a traditional ITS (i.e., no game features), subsequent performance during game and nongame practice methods was equivalent. Additionally, previous work with the game-based aspects in iSTART–ME has shown consistent positive effects for motivation and enjoyment (Jackson, Davis, et al., 2011; Jackson & McNamara, 2011). This combination of results leads to two hypotheses regarding the current study.

One hypothesis is that games improve motivation and enjoyment, but they may impede learning, especially initially (Adams, Mayer, MacNamara, Koenig, & Wainess, 2012; Jackson et al., 2012). In this case, we would expect the game-based environment to produce lower learning outcomes than the traditional ITS, particularly in the initial stages of learning. The second hypothesis is that the game-based components of iSTART–ME improve motivation and enjoyment (Cordova & Lepper, 1996; Papastergiou, 2009), and this increase in affective measures mediates learning (Alexander et al., 1997). This hypothesis suggests that students in the game-based training should see improved motivation and enjoyment over time and should see a corresponding increase in performance during the later stages of training (compared with the traditional ITS).

Procedure

Participants and setting. Eighty-four high school students were recruited from the general city-wide high school population in an urban environment in the mid South (51% male; 81% African American, 13% White, 6% other; average grade completed = 10th grade; average age = 15.8 years). The 11-session experiment was conducted in a research laboratory on a large university campus and involved four phases: pretest, training, posttest, and retention test.

Pretest. During the first session, students completed a pretest that included questions to collect basic demographics, prior motivation (including selected questions adapted from the Motivated Strategies for Learning Questionnaire, or MSLQ; Pintrich, Smith, Garcia, & Mckeachie, 1993), and an assessment of their prior ability to self-explain (described in more detail later).
Training. At beginning of the eight training sessions, participants completed a 12-item daily survey (see Measures section for details). After the daily survey, students then interacted with their randomly assigned between-subjects condition: a game-based system (iSTART-ME, n = 41) or a traditional ITS (iSTART-Regular, n = 43). Students in the educational game condition interacted with the full game-based selection menu in iSTART-ME across eight separate sessions of at least 1 hr 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 hr each).

The initial training within both conditions 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. Participants assigned to game-based training were then free to use the full selection menu (Figure 2), while the ITS students continually transitioned from one text to another within the Coached Practice environment (Figure 1).

Like many ITSs, iSTART–Regular is not completely void of mechanisms and features that are commonly used within games. For example, iSTART–Regular displayed points for each self-explanation (near bottom-left of Figure 1), included adaptive feedback from an animated agent and provided a trophy (or lack thereof) based on the performance within each text. These features (points, personalized feedback, animated characters, and trophies/badges) are commonly used in numerous types of games and systems, both virtual and physical. Table 1 provides a more thorough comparison of the two training systems in terms of the key features included in each. iSTART–ME differed from iSTART–Regular primarily in the presence of the selection menu, which allowed participants to play mini-games and modify certain aspects of the environment (e.g., swap tutors, personalize their avatar). Both systems allowed students to progress through the tutoring at their own pace, and therefore, not all students experienced the same components at the same time. This is a key characteristic of ITSs and virtually all games that adapt interactions on the basis of user decisions. Hence, some students naturally receive more or different kinds of training and practice than others.

Posttest and retention. All students completed the posttest and then a delayed retention test (completed a week after posttest). The posttest consisted of assessments similar to those from the pretest (details are discussed in the Measures section). These included measures of self-explanation ability and students’ motivation during the study, along with questions pertaining to students’ attitudes, perceptions, and experiences. The retention test was used to assess the durability of students’ self-explanation skills after a 1-week delay without training.

Measures

Survey and performance measures were collected during pretest, training, posttest, and retention. These included measures related to self-explanation ability as well as students’ attitudes, motivation, self-efficacy, and enjoyment.

Self-explanation ability. Students’ performance on self-explanation tasks was collected during pretest, training, posttest, and retention. During training, students interacted with various texts and all self-explanations were scored through the iSTART assessment algorithm and recorded into a database. During each of the three testing phases, students were presented with a new text (not included within training) and prompted to self-explain specific sentences (eight self-explanations during each test). These three texts were selected due to their similarity in terms of length (281–329 words), content difficulty (Grade Level 8–9), and linguistic features (i.e., similar scores on the five principal component scores within Coh-Metrix; Graesser, McNamara, & Kulikowich, 2011). Each self-explanation provided by the students was scored using the iSTART assessment algorithm, the performance of which has been shown to be comparable to that of human scorers (Jackson, Guess, et al., 2010). Unfortunately, due to a technical error, the three texts were not automatically counterbalanced across the testing phases. Thus, despite extensive efforts to utilize equitable texts, comparisons of self-explanations across time should be interpreted with caution and must be replicated using appropriate methodology. Nonetheless, the lack of counterbalancing should not affect any comparisons between conditions.

Attitudes, motivation, self-efficacy, and enjoyment. Survey questions were included during pretest, posttest, and daily training sessions to assess students’ attitudes, motivation, self-efficacy, and enjoyment. Pretest and posttest measures included several questions adapted from the MSLQ (Pintrich et al., 1993). The questions adapted from the MSLQ were selected to address students’ motivation and self-efficacy. In addition to these standardized measures, questions were included from previous research with the iSTART system (Jackson, Davis, et al., 2011; Jackson, Graesser, & McNamara, 2009; Jackson & McNamara, 2011). These additional questions were implemented within the pretest, posttest, and daily surveys and were designed to measure students’ self-assessments of motivation, expectations for system interactions, current affect and mood, and overall enjoyment of the system.

Table 1

<table>
<thead>
<tr>
<th>Game Mechanism and Feature Differences Between iSTART–ME and iSTART–Regular</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mechanisms</strong></td>
</tr>
<tr>
<td>Feedback</td>
</tr>
<tr>
<td>Incentives</td>
</tr>
<tr>
<td>Control</td>
</tr>
<tr>
<td>Task difficulty</td>
</tr>
<tr>
<td>Environment</td>
</tr>
</tbody>
</table>
The daily surveys used within the current study have been used in previous research with the game-based version of iSTART–ME (Jackson, Davis, et al., 2011; Jackson & McNamara, 2011). These surveys were designed to assess students’ moods, attitudes, and perceptions across the time frame of an experiment without being an invasive measurement during system interactions. These surveys were administered at the beginning of each training session and specifically addressed students’ experiences from the previous session (overall impression, enjoyment, boredom, frustration, interface problems, feeling of learning, feeling of improvement) and assessed their current attitudes and feelings (current mood, anticipation about participating, level of motivation, intention to perform well, desire to do better than others).

Results

Training Sessions

As mentioned previously, both systems allowed students to progress through training at their own speed. Despite the adaptivity and self-paced interactions, students’ prior ability was not related to the amount of practice students received during training. More specifically, self-explanation ability at pretest was not related to the number of practice texts that students completed ($r = .079, p = .45$). Therefore, initial ability levels were not related to the amount of extended practice that students received, and most students experienced the training components at approximately the same time. The vast majority of students completed the two regular practice texts and transitioned into the extended practice during the first ($n = 6$) or second ($n = 72$) session, while some students did not reach the extended practice section until the third ($n = 5$) or fourth ($n = 1$) session. Ultimately, all students completed the training modules and subsequently interacted with their randomly assigned training condition for the remainder of the study.

Attitudes, Motivation, Self-Efficacy, and Enjoyment

User experience measures from pretest questions, daily surveys, and posttest questions were analyzed to explore students’ attitudes, perceptions, and experiences within the two training systems. Analyses on the pretest survey questions indicate that there were no significant differences between conditions on questions prior to the start of training that related to enjoyment, motivation, self-efficacy, or competitiveness (see Table 2).

The posttest survey included several questions related to enjoyment, perceived learning, and usability within the system (see Table 2 for descriptive and analysis of variance [ANOVA] results). A posttest enjoyment and motivation composite score was created by averaging across six separate questions. An ANOVA on the enjoyment and motivation composite score yielded a significant effect of condition, $F(1, 82) = 8.28, p = .005$, mean square error ($MSE$) = 1.15, Cohen’s $d = 0.628$. These results indicate that the game-based environment was rated as a significantly more positive experience than the traditional ITS. Additionally, a composite scale that assessed students’ perceived learning, effort, and values for the target system and materials found no significant differences between the game and nongame system. Likewise, a four-question scale that assessed system usage and interface confusion revealed no significant differences between conditions. These results suggest that the game-based selection menu system was more enjoyable and motivating, but just as valuable and easy to use as the ITS.

Daily surveys were administered to assess students’ reports of their previous-session experiences and current-session expectations. Questions related to similar concepts were combined into several composite scores (i.e., enjoyment during the previous session, improvements in self-efficacy, and motivation for the current session). A composite score was created for enjoyment during the previous session by combining scores from the following three questions: “My most recent session was . . . (very bad = 1, very good = 6),” “I enjoyed my most recent session . . . (not at all = 1, very much = 6),” and “I was bored during my most recent session . . . (reversed scored; all the time = 1, never = 6).” A mixed-factor ANOVA on this composite score indicated that there was a significant main effect for condition, such that students in the game-based condition rated their session experiences more favorably ($M = 4.89$, standard error [SE] = 0.159) than did students in the ITS condition ($M = 4.07$, SE = 0.151), $F(1, 76) = 13.92, p < .001, MSE = 7.51$. There was also a significant linear interaction between session and condition, $F(1, 76) = 3.266, p = .004, MSE = 0.606$ (see Figure 5). Pairwise comparisons using Bonferroni adjustments for multiple tests confirmed that enjoy-

---

Table 2

<table>
<thead>
<tr>
<th>Survey scales</th>
<th>iSTART–ME</th>
<th>iSTART–Regular</th>
<th>$F(1, 82)$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected Enjoyment and Motivation to Participate (5 items)</td>
<td>4.73 (0.66)</td>
<td>4.69 (0.73)</td>
<td>0.07</td>
<td>.799</td>
</tr>
<tr>
<td>Achievement Motivation and Learning Values$^a$ (7 items)</td>
<td>5.31 (0.98)</td>
<td>5.31 (0.86)</td>
<td>0.00</td>
<td>.996</td>
</tr>
<tr>
<td>Self-Efficacy (3 items)</td>
<td>5.94 (1.12)</td>
<td>5.95 (0.91)</td>
<td>0.00</td>
<td>.962</td>
</tr>
<tr>
<td>Competitiveness (2 items)</td>
<td>4.91 (1.08)</td>
<td>4.84 (1.21)</td>
<td>0.10</td>
<td>.758</td>
</tr>
</tbody>
</table>

Posttest survey measures

| Enjoyment and Motivation (6 items) | 4.55 (1.09) | 3.83 (1.20) | 8.28 | .005 |
| Perceived Learning, Effort, and Values$^a$ (9 items) | 4.81 (1.56) | 4.53 (1.61) | 0.64 | .425 |
| Ease of Use (4 items) | 3.32 (1.04) | 3.31 (1.20) | 0.00 | .951 |

$^a$Questions adapted from the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich, Smith, Garcia, & Mckeachie, 1993).

---

1 The mixed-factor ANOVA results presented here are based on the linear equation contrasts for the interaction. The overall within-subject interaction effects for this mixed-factor ANOVA (including Huynh–Feldt corrections due to significant sphericity and a large Greenhouse–Geisser $\varepsilon > .75$) were also significant, $F(5.99, 455.26) = 3.27, p = .004$. 
ment tended to increase across sessions for students interacting with the game-based system and decrease for those in the ITS condition. Specifically, for students in the iSTART–ME condition, overall enjoyment at Sessions 5 ($t = 4.18, p_{adj} < .01$) and 8 ($t = 5.03, p_{adj} < .001$) were significantly higher than at Session 1 (with the middle sessions being roughly equivalent). By contrast, students in the ITS condition provided their highest ratings during the first two sessions, with enjoyment at Session 2 being significantly higher than at Session 4 ($t = 3.40, p_{adj} < .05$).

Similarly, a composite score was created for the daily survey questions related to students’ improvements in self-efficacy. This score combined student ratings for two questions about the previous session, “I felt like I learned the material . . . (not at all = 1, very much = 6)” and “I feel like my reading skills improved . . . (not at all = 1, very much = 6).” A mixed-factor ANOVA on the self-efficacy composite score did not indicate a significant main effect for condition, $F(1, 76) = 2.50, p = .118, MSE = 7.43$, but did reveal a significant linear interaction between session and condition, $F(1, 76) = 2.91, p = .015, MSE = 0.673$ (see Figure 6). This interaction reflects the finding that students’ reported self-efficacy increased across sessions if they had interacted with the game-based version of training and decreased if they interacted with the traditional ITS. Specifically, pairwise comparisons (using Bonferroni adjustments) showed that iSTART–ME students provided their highest self-efficacy rating in the final session (Session 8 was marginally higher than Session 4, $t = 3.15, p_{adj} = .09$). In contrast, iSTART–Regular students provided their highest self-efficacy ratings in the first two sessions (Session 1 was significantly higher than Session 7, $t = 3.35, p_{adj} < .05$; and Session 2 was significantly higher than both Session 4, $t = 3.57, p_{adj} < .05$, and Session 7, $t = 3.52, p_{adj} < .05$).

Finally, a composite score was created for the daily survey questions that pertained to motivation to participate in current session: “My mood right now is . . . (very negative = 1, very positive = 6),” “I am looking forward to participating in today’s session . . . (not at all = 1, very much = 6),” “I am motivated to participate in today’s session . . . (not at all = 1, very much = 6),” and “I plan to do my best during today’s session . . . (not at all = 1, very much = 6).” A mixed-factor ANOVA yielded a marginal main effect of condition, $F(1, 75) = 3.05, p = .085, MSE = 5.82$, indicating that students in the game-based condition ($M = 5.30, SE = 0.142$) tended to be more motivated to participate than students interacting with the ITS ($M = 4.96, SE = 0.133$). This mixed-factor ANOVA also revealed a significant linear interaction between session and condition, $F(1, 75) = 4.95, p = .029, MSE = 0.410$ (see Figure 7), reflecting the finding that students’ motivation to participate in the current session remained stable for those in the iSTART–ME condition but declined in the iSTART–Regular condition. Pairwise comparisons (using Bonferroni adjustments) confirmed that students’ ratings for today’s session within the game-based system were not significantly different across sessions ($p_{adj} > .05$), and students within the ITS provided marginally higher ratings in Sessions 1 and 2, compared with Sessions 4 ($t_{Session1} = 3.24, t_{Session2} = 3.26, p_{adj} < .10$) and 7 ($t_{Session2} = 3.15, p_{adj} < .10$).

These results collectively indicate that students provided equivalent ratings in the two conditions for the first two sessions (when training was the most similar), but after the game-based aspects

2 The mixed-factor ANOVA results presented here are based on the linear equation contrasts for the interaction. The overall within-subject interaction effects for this mixed-factor ANOVA (including Greenhouse–Geisser corrections due to significant sphericity and a moderate Greenhouse–Geisser $\varepsilon < .70$) were also significant, $F(4.75, 360.84) = 2.91, p = .015$.

3 The mixed-factor ANOVA results presented here are based on the linear equation contrasts for the interaction. The overall within-subject interaction effects for this mixed-factor ANOVA (including Huynh–Feldt corrections due to significant sphericity and a large Greenhouse–Geisser $\varepsilon > .80$) were also significant, $F(6.22, 466.40) = 2.09, p = .050$.

Figure 5. Composite means for enjoyment questions about the previous session. ITS = intelligent tutoring system.

Figure 6. Composite means for self-efficacy daily survey questions. ITS = intelligent tutoring system.

Figure 7. Composite scores for questions about students’ motivation to participate in the current session. ITS = intelligent tutoring system.
were made available, students interacting with the educational games provided more positive ratings than did students interacting with the ITS. In sum, the combined evidence from the daily surveys and posttest questions indicates that students preferred to interact with the game-based system more so than the traditional tutoring system.

**Learning Outcomes**

Analyses were conducted on the self-explanation scores from the pretest, posttest, and retention test. All self-explanations were scored using the iSTART assessment algorithm which has high correspondence to human scores ($\kappa = .646$; Jackson, Guess, et al., 2010; McNamara, Boonthum, Levinstein, & Millis, 2007). As shown in Figure 8, self-explanation quality improved from pretest to posttest for students in both conditions, and this increase in performance was maintained in a delayed retention test 1 week later, but there was no benefit for either condition. Specifically, a mixed-factor ANOVA confirmed a main effect of test, $F(1, 82) = 22.67, p < .001, MSE = 0.20$, reflecting the finding that self-explanation quality scores did not differ from posttest to retention test ($t < 1$), but both the posttest ($t = 7.19, p < .001$) and retention test ($t = 7.77, p < .001$) were significantly higher than the pretest (see Figure 8). There was no effect of condition, $F(1, 82) = 1.61, p = .21, MSE = 0.71$, and no interaction between condition and test, $F(1, 82) = 0.48, p = .49, MSE = 0.17$.

One of the limitations of this study is that the self-explanation texts were not counterbalanced among the pretest, posttest and retention test phases. Therefore, the pretest to posttest improvement in self-explanation ability is conflated with a potential text effect. However, combining these findings with the improvement of self-explanation ability during training lends support to students’ improvement between testing phases. Additionally, the lack of counterbalancing does not affect the comparisons between conditions at each phase. Thus, the equivalent performance between conditions at each testing phase is not confounded by text.

We also conducted analyses to examine self-explanation performance comparing conditions during extended training. The first training sessions included the complete Introduction and Demonstration modules, along with the first two texts in regular practice (initial ~2 hr training). On average, students began interacting with the two different extended practice modules during the second session (i.e., students started using either only coached practice or the full selection menu during the second session). A mixed-factor ANOVA on the frequency of self-explanations yielded significant main effects for both session, $F(1, 67) = 20.15, p < .01, MSE = 35.71$, and condition, $F(1, 67) = 4.03, p < .05, MSE = 384.86$, but revealed a nonsignificant interaction between session and condition, $F(1, 67) = 1.13, p = .29, MSE = 46.18$ (see Figure 9 for mean frequencies across days). Students who interacted with the ITS produced more self-explanations ($M = 23.39, SE = 1.14$) during extended practice than did students within the game-based training ($M = 20.03, SE = 1.23$), and the number of self-explanations tended to be highest in Session 2. Pairwise comparisons (using Bonferroni adjustments) indicate that students across conditions generated an equivalent number of self-explanations during regular practice (i.e., Session 1) and that the frequency of self-explanations during the first session was significantly lower than in all other sessions (all $t$s > 6.60, all $p$s < .05). For training that took place within extended practice (i.e., Sessions 2–8), participants within the ITS produced significantly more self-explanations than students using the game-based system during Sessions 3 and 5 ($t_{\text{Session}3} = 2.16, p_{\text{adjusted}} = .035$; $t_{\text{Session}5} = 2.04, p_{\text{adjusted}} = .046$) and were marginally higher for Sessions 2 and 8 ($t_{\text{Session}2} = 1.97, p_{\text{adjusted}} = .053$; $t_{\text{Session}8} = 1.91, p_{\text{adjusted}} = .060$).

Further analyses examined students’ self-explanation quality as computed by the iSTART assessment algorithm across the eight training sessions (see Figure 10). The two main hypotheses in the current study predicted opposite slopes for game-based performance during the initial and later training sessions. Specifically, the first hypothesis predicted a negative slope for game-based performance during the early sessions, while the second hypothesis predicted a positive slope for game-based performance in later sessions (this predicted decrease followed by an increase was tested through a quadratic contrast). A mixed-factors ANOVA did not indicate a significant main effect for condition, $F(1, 67) = .085$.

---

4The mixed-factor ANOVA results presented here are based on the linear equation contrasts for the interaction. The overall within-subject interaction effects for this mixed-factor ANOVA (including Greenhouse–Geisser corrections due to significant sphericity and a moderate Greenhouse–Geisser $\varepsilon < .70$) were marginally significant, $F(4.62, 309.69) = 1.994, p = .085$. 

---

![Figure 8](image_url)  
**Figure 8.** Self-explanation performance during testing phases (means and standard errors). ITS = intelligent tutoring system.

![Figure 9](image_url)  
**Figure 9.** Frequency of self-explanations across sessions. ITS = intelligent tutoring system.
students do not enjoy the experience, they are likely to cease or avoid further interactions, which is particularly detrimental to systems that require skill development over longer periods of training. These long-term skill acquisition systems must be designed to foster significant increases in mastery development, but they must also be enjoyable to use. In the case of strategy tutors, these systems must not only teach the strategies themselves but provide an effective, motivating practice environment where students can apply this training and sufficiently develop the target skills into more automatic and stable processes. For example, previous research with iSTART has illustrated the need for prolonged training of at least 5 days with the system, such that students (specifically those with low prior abilities) have sufficient opportunities to apply and master the target skills (Jackson, Boonthum, et al., 2010). Thus, a game-based version of the system (iSTART-ME) was designed to maintain higher levels of student motivation and engagement over an extended practice period by incorporating and leveraging mechanisms that positively influence affect (Conati, 2002; Corbett & Anderson, 2001; Cordova & Lepper, 1996; Graesser et al., 2009; Malone & Lepper, 1987; Moreno & Mayer, 2005; Shute, 2008). The current work focused on evaluating the global impacts of this game-based learning environment and comparing it to a traditional ITS. Additionally, in this study we investigated the specific time-based effects of these systems on both motivational and learning outcomes.

Within the current study, the game-based version of training was preferred significantly more than the traditional tutoring system. The results from the posttest survey indicate that students perceived both systems to be equally helpful and easy to use but that the game-based system was significantly more motivating and enjoyable (Table 2). Likewise, results from the daily surveys (Figures 5–7) illustrate that students who interacted with the game-based system tended to improve in their perception of the system across sessions, have improved self-efficacy (compared with those interacting with the ITS), and slowly increase (or at least maintain) motivation for future interactions. In contrast, daily ratings by students who interacted with the traditional tutoring system decreased in enjoyment, motivation, self-efficacy, and desire for future interactions. The game components present within iSTART-ME seem to be activating related motivational constructs that remain effective across time. These trends are also fairly gradual, indicating that changes may occur in smaller increments and slowly build up with more iterative interactions (possibly suggesting a cycle of affective improvement across time).

The results for self-explanation performance (the targeted skill) provide a more complex message. The self-explanation frequencies and means (see Figures 9 and 10) help to provide significant insight into the learning trajectories comparing game-based and traditional tutoring systems. Students within game-based training generated significantly fewer self-explanations than students using the ITS. This difference is likely due to time spent with the additional nongenerative activities available within the game-based selection menu (i.e., mini-games, personalizing their avatar).

### Discussion

The goal of tutoring systems and educational games is to produce effective and enjoyable learning experiences. However, if

---

**Figure 10.** Self-explanation performance during training (means and standard error). ITS = intelligent tutoring system.

2.89, \( p > .05, \) MSE = 1.27; however, there was a significant quadratic interaction between session and condition, \( F(1, 67) = 22.91, p < .001, \) MSE = 0.11. Students’ self-explanation quality in the iSTART-ME condition tended to decrease during the initial interactions with the game-based selection menu and educational games (Sessions 3, 4, and 5), and Bonferroni-adjusted pairwise comparisons indicated that scores were significantly different between the two training conditions for Sessions 3 (\( t = -2.05, p_{\text{adjusted}} < .05 \)), 5 (\( t = -2.77, p_{\text{adjusted}} < .01 \)), and 6 (\( t = -2.21, p_{\text{adjusted}} < .05 \)). These results are partially attributable to the reduction in direct feedback from Merlin in Coached Practice. These trends may also be partially due to the additional cognitive tasks involved with learning the menu itself, its features, and various game dynamics, in addition to the targeted self-explanation strategies. Indeed, the analyses related to Figure 9 demonstrate that students within the game-based system produced fewer self-explanations and thus practiced less on the target task. Despite these extra features and time spent off-task (i.e., not practicing), the students within the game-based system were able to compensate for the initial deficit over time and ultimately rose to match the performance of the ITS participants. It is important to note that the students within the game-based training were more motivated to participate, enjoyed interacting with the system more, and had larger improvements in self-efficacy than those students in the ITS condition, which would be a crucial factor in a real-life situation such as a classroom or practicing at home.

These results concur with findings in two studies conducted by Jackson, Dempsey, et al. (2011, 2012), collectively suggesting that game elements have the potential to detract from learning during initial skill acquisition. However, game environments can provide a more positive experience over time. Thus, the game-based system investigated in this study appears to strike an appropriate balance between both learning and enjoyment, improving on the imbalance previously encountered within a traditional tutoring system. This finding is especially encouraging for strategy-based tutors that require long-term interactions for students to develop skill mastery.

---

\[5\] The mixed-factor ANOVA results presented here are based on the contrasts for a quadratic interaction. The overall within-subject interaction effects for this mixed-factor ANOVA (including Greenhouse–Geisser corrections due to significant sphericity and a moderate Greenhouse–Geisser \( \epsilon < .75 \)) were also significant, \( F(5.00, 335.18) = 4.23, p = .001. \]
selecting a character, changing the interface colors, and so on). Despite the increased number of self-explanations for the ITS condition, both training systems produced equivalent self-explanation performance at the posttest and delayed retention test (Figure 8).

Based on the analyses for Figure 10, it appears that the traditional ITS system showed a predominantly positive relation between the amount of training and performance. This trend was expected, based on past research substantiating the positive benefits of iSTART training over time (Jackson, Boonthum, et al., 2010; McNamara, O’Reilly, Rowe, Boonthum, & Levinstein, 2007). The trajectories for self-explanation quality within the game-based system allow us to address the primary hypotheses. Our two main hypotheses regarded the potential benefits or hindrances from the game-based version of training. Specifically, the first hypothesis predicted that the addition of game-based features may detract from the learning objectives, such that students should exhibit a decrease in performance during the initial stages of training. Indeed, the decrease in performance during Sessions 3 through 5 (Figure 10) suggest that the game-based features may initially detract or interfere with students’ ability to apply the target strategies (possibly due to competing stimuli and accommodating multiple goals).

The second hypothesis predicted that game-based features should improve motivation and engagement during prolonged periods of training, which should have a corresponding increase in applied mastery (i.e., increased performance during later practice). The increase in self-explanation quality across Sessions 6 through 8 (Figure 10) lends support for this second hypothesis. Specifically, the increase in performance during these sessions corresponds with the improved affect and motivation ratings in the game-based condition (Figures 5–7).

Analyses on the self-explanation quality across time indicated that the game-based system resulted in a significant quadratic relation between training and performance, such that performance initially declined and subsequently increased. This curvilinear performance trajectory provides statistical support for both hypotheses and may also help to explain some of the mixed results found in the previous literature on educational games. Specifically, the time scale of measurement within a study may determine whether performance trends for game-based systems appear to be positive, negative, or neutral.

It is also worth noting that the minimal game features remaining in the ITS (see Table 1) were not enough to produce the same motivational improvements as the fully game-based version of training. This finding is potentially significant for two reasons. First, the fully game-based training likely would have produced even larger motivational differences if it had been compared with a more stripped-down version of an ITS (i.e., exaggerating the already significant effects). Second, just adding in a few game-like features to an ITS is not enough to produce the effects found in more coherent and contextually bound educational games. Our findings in this study demonstrate that the combined set of features and mechanisms integrated within our game-based system (feedback, incentives, control, task difficulty, and environment) effectively enhanced users’ experience with the tutoring system and that most of these benefits tended to remain stable or even increase across time. The overall findings indicate that game-based interaction mechanisms can provide enjoyable, effective interactions that promote sustained motivation and mastery over time.

The current results further suggest that future research on educational games should incorporate multiple time scales of measurement to investigate the complex trajectories of both learning and motivation within these environments. As suggested from the current work, along with results from Jackson, Dempsey, et al. (2011, 2012), isolated measurements solely at pretest and posttest may provide an oversimplified snapshot of the potential benefits (or weaknesses) of game-based education. The current work is intended to inform researchers’ future development and evaluation, as well as contribute to the need for empirical comparisons between game-based and nongame-based tutoring environments (O’Neill & Fisher, 2004, O’Neill et al., 2005).

The outcomes and concepts discussed here provide unique insight into various time-based effects within educational games. Repeated observations allow us to represent students’ experiences throughout the interaction process and are further supported with more summative measures collected separately from training (i.e., posttest and delayed retention). The results from this study as well as our previous studies (Jackson, Dempsey, et al., 2011, 2012) support the assumption that students prefer working with game-based tutoring environments and that, over time, these systems can provide enjoyable training that produces learning outcomes comparable to more traditional ITSs. The current work provides substantial support for incorporating games into long-term tutoring environments and should help researchers and educators to better understand the potential benefits from these game-based components and systems.

References


Jackson, G. T., Graesser, A. C., & McNamara, D. S. (2009). What students expect may have more impact than what they know or feel. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. C. Graesser (Eds.), *Artificial intelligence in education: Building learning systems that care. From knowledge representation to affective modeling* (pp. 73–80). Amsterdam, the Netherlands: IOS Press.

