Natural Language Processing and Game-based Practice in iSTART

G. Tanner Jackson[1], Chutima Boonthum-Denecke[2], and Danielle S. McNamara[1]

[1]Learning Sciences Institute, Arizona State University
[2]Hampton University

Jackson, T., Boonthum-Denecke, C., & McNamara, D. (2015). Natural language processing and game-based practice in iSTART. Journal of Interactive Learning Research, 26(2), 189-208.

Author's Note

This work was supported by the Institute of Education Sciences (IES) through grant R305A130124. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily represent views of the IES.

Natural Language Processing and Game-based Practice in iSTART

G. Tanner Jackson[1], Chutima Boonthum-Denecke[2], and Danielle S. McNamara[1]

[1]Learning Sciences Institute, Arizona State University

[2]Hampton University

Abstract

Intelligent Tutoring Systems (ITSs) are situated in a potential struggle between effective pedagogy and system enjoyment and engagement. iSTART, a reading strategy tutoring system in which students practice generating self-explanations and using reading strategies, employs two devices to engage the user. The first is natural language processing (NLP). Incorporating NLP within iSTART allows students to use their own thoughts and ideas to communicate with the system, and serves as the core *intelligence* of the system that is used to drive the feedback and the adaptive interactions during practice. Studies have shown that the NLP algorithms within iSTART perform comparably to human raters and provide a good measure for the sophistication of student self-explanations. The second device is the use of game-based practice. Skill mastery requires a significant commitment to practice over extended periods of time. Unfortunately, this persistent and repetitive practice is also associated with disengagement from the target educational task. Therefore, a gaming environment was developed that integrates multiple combinations of enjoyable, engaging game elements with the target practice tasks. This paper describes these two principle aspects of iSTART and research on their effectiveness.

Natural Language Processing and Game-based Practice in iSTART

iSTART (Interactive Strategy Training for Active Reading and Thinking) is an Intelligent Tutoring System (ITS) designed to improve high school and college students' reading comprehension by providing instruction on how to self-explain using effective reading strategies. iSTART first introduces students to the concept of self-explanation and then provides instruction on how to use reading comprehension strategies such as paraphrasing, generating bridging inferences, and elaboration to improve self-explanations and ultimately comprehension (e.g., Magliano et al., 2005; McNamara, 2004; McNamara, O'Reilly, Best, & Ozuru, 2006). After they are introduced to the strategies and given examples of how they are used, students then practice generating self-explanations while reading science texts.

iSTART was originally modeled after a human-based intervention called Self-Explanation Reading Training, or SERT (McNamara, 2004; McNamara & Scott, 1999; O'Reilly, Best, & McNamara, 2004). The automated iSTART system produces gains equivalent to the human-based SERT program (O'Reilly, Sinclair, & McNamara, 2004; O'Reilly, Best, & McNamara, 2004). Both the live and the automated interventions included an introduction to the strategies, demonstration of their use, and practice using them while reading science texts. They both implemented the pedagogical principle of modeling-scaffolding-fading across the introduction, demonstration, and practice phases. Nonetheless, there are several key differences between iSTART and SERT. First, unlike SERT, iSTART is web-based and thus accessible at a distance. Second, it is automated, and hence it can work with students on an individual level and provide self-paced instruction; the student can stop and start at any time. Third, rather than a teacher or instructor, iSTART incorporates animated agents that engage students with the system and tutor them on how to correctly apply various reading strategies. The agents were designed to introduce students to the concept of self-explanation and to demonstrate specific strategies to enhance their reading comprehension. For example, the introduction module uses a classroom-like discussion format between three agents (a teacher and two student agents) to present the relevant reading strategies within iSTART. These agents interact with each other, providing students with information, posing questions to each other, and giving example explanations to illustrate appropriate strategy use (including counterexamples). These interactions exemplify the active processing that students should use when providing their own self-explanations. A fourth difference between the two interventions is that, by necessity, iSTART uses natural language processing (NLP) to interpret the students' self-explanations and subsequently provide feedback.

Insert Figure 1

iSTART NLP Algorithm

As illustrated in Figure 1, feedback provided in iSTART is driven by NLP algorithms. First the student enters a response, which in the case of iSTART is an explanation of the sentence or multiple sentences in a text. The response is in the form of natural language. That is, the student does not choose a response from list of preset responses or choices. The response is open ended and potentially ungrammatical, ambiguous, ill-formed, and ridden with spelling errors (Renner, McCarthy, Boonthum-Denecke, & McNamara, 2011). The algorithm drives the feedback that is given to the student. As such, the algorithm and the feedback together comprise the heart of the *intelligence* in iSTART. The NLP algorithms provide the means for the system to respond to the student and adapt the training to the student's needs. For example, once the students are in the practice module, an animated character (Merlin) provides feedback on students' explanations, prompting them to generate new explanations using their newly acquired repertoire of strategies (see Figure 2 for screenshot). The main focus of the practice module is to provide students with an opportunity to apply the reading strategies to new texts and to integrate their knowledge from different sources in order to understand a challenging text. Their explanation may include world and domain knowledge or it may stem from prior sentences in the text. Merlin provides feedback for each explanation generated by the student. For example, he may prompt them to expand the explanation, ask the students to incorporate more information, or suggest that they make a connection back to other parts of the text. The iSTART algorithm is designed to assess the quality of the student's response such that it can drive Merlin's feedback to the student in pedagogically effective ways.

Insert Figure 2

The iSTART assessment algorithm evaluates each student self-explanation as a 0, 1, 2, or 3 (see Table 1 for examples). An assessment of "0" relates to explanations that are either too short or contain mostly irrelevant information. An iSTART score of "1" is associated with an explanation that primarily relates only to the target sentence itself (sentence-based). A "2" means that the student's explanation incorporated some aspect of the text beyond the target sentence (text-based). If an explanation earns a "3" from the iSTART evaluation, then the explanation incorporates information at a global level, and may include outside information or refer to an overall theme across the whole text (*i.e.*, global-based information).

Insert Table 1

Determining the appropriate feedback for each explanation depends on the accuracy of the evaluation algorithm implemented within iSTART. Obviously the feedback has the potential to be more appropriate when the evaluation algorithm more accurately depicts explanation quality and related characteristics. In order to accomplish this task and interact with students in a meaningful way, the system must be able to adequately interpret natural language text explanations.

Several versions of the iSTART evaluation algorithm have been tested and validated with human performance (McNamara, Boonthum, Levinstein, & Millis, 2007). The resulting algorithm utilizes a combination of both word-based approaches and latent semantic analysis (LSA; Landauer, McNamara, Dennis, & Kintsch, 2007). The word-based approaches provide a more accurate picture of the lower level explanations (ones that are irrelevant, or simply repeat the target sentence). They are able to provide a finer distinction between these groups than does LSA. In contrast, LSA provides a more informative measure for the higher level and more complex explanations. Therefore, a combination of these approaches is used to calculate the final system evaluation.

The word-based approach originally required a significant amount of hand-coded data, but now uses automatic methods when new texts are added. The original algorithm required experts to create a list of "important" words for each text and then also a list of associated words for each "important" word. This methodology was replaced, and now the word-based component relies on a list of content words (nouns, verbs, adjectives, adverbs) that are automatically identified in the text (McNamara et al., 2007). The word-based assessment also includes a length criterion where the student's explanation must exceed a certain number of words (calculated by multiplying the number of words in the target sentence by a pre-specified coefficient).

The LSA-based approach uses a set of benchmarks to compare student explanations to various text features. These LSA benchmarks include 1) the title of the passage, 2) the words in the target sentence, and 3) the words in the previous two sentences. The third benchmark originally involved only words from causally related sentences, but this required conducting a discourse analysis of each text, and thus was replaced by the words from adjacent sentences. Within the science genre, this replacement was expected to do well, because of the linear argumentation most often employed in science textbooks. However, it has not been established how well these assessment metrics apply to texts from other domains.

The evaluation of the iSTART algorithm was originally conducted using only a few practice texts within iSTART (McNamara et al., 2007). Students self-explained target sentences within a text, and those self-explanations were assessed separately by the iSTART algorithm and by human raters. Figure 3 displays the agreement between the scores from the iSTART algorithm compared to the human scores. These studies showed that there was a high correspondence (r = 0.64 - 0.71; perfect agreement = 62 - 64%), particularly at the extremes, such as when humans rated a self-explanation as globally focused and high quality (i.e., a 3) and iSTART gave the explanation a 3 or both humans and the algorithm scored the self-explanation as being poor (i.e., a 0 or 1), with d-primes all above 1.5.

Insert Figure 3

Subsequent studies have been conducted that evaluated the assessment performance on a variety of untrained texts which were added to the system after the algorithm had been developed and implemented in vivo (Jackson, Guess, & McNamara, 2010). These studies included a set of 5,400 student self-explanations collected within iSTART from a variety of science texts. Each self-explanation was rated by three trained human judges. These human raters were extensively trained on self-explanation strategies, had little or no knowledge in how the system algorithm worked, and the ratings were provided independently of the iSTART algorithm scores (i.e., raters never saw the output from iSTART). Inter-rater reliability on a training set of data (including all 3 raters) resulted in an average correlation of .70. Figure 4 displays the agreement rating between iSTART and humans for the untrained texts (kappa = 0.646; Jackson, Guess, & McNamara, 2010). Across the evaluations of the iSTART algorithm in both studies (i.e., Figures 3 and 4), it appears that humans and iSTART agree on explanations that are nonsensical or irrelevant (both rate as a score of 0), sentence-based (both rate as a score of 1), and global-based (both rate as a score of 3). The studies further indicate that the text-based explanations (score of 2) are more difficult to distinguish. This has been true for both humans and the iSTART algorithms.

Insert Figure 4

These results suggest that the iSTART algorithm has the ability to adapt to new texts and information in an appropriate and informative manner. The results also indicate that iSTART's evaluations are sufficiently accurate compared to humans, and can provide a general indicator for the amount of processing required to generate self-explanations (i.e., the degree to which students are processing the information at the sentence, text, or global level).

Student Performance in iSTART

Evaluations have shown that iSTART accurately assesses student self-explanations, and therefore has the ability to provide students with tailored feedback. Several laboratory studies have confirmed that iSTART improves students' ability to self-explain and their ability to comprehend challenging texts (e.g., Magliano et al., 2005; McNamara et al., 2006). More recently, we reported the results from a long-term, ecological study to assess the degree to which iSTART helped to improve students' self-explanation quality in a classroom setting (Jackson, Boonthum, & McNamara, 2010). Participants in this study included 389 high school students in science classes. Throughout the course of an academic year, students spent time each week interacting with iSTART after having completed the initial training (introduction module, demonstration module, and initial practice module). During this extended practice phase, students self-explained texts from the iSTART library as well as texts assigned by their teachers. Students interacted with the system at their own pace and therefore experienced a different number of total texts. As shown in Figure 5, results from this school-based study (partially reported in Jackson, Boonthum, & McNamara, 2010) confirmed that students improved performance over time. Learning curves and regression analyses confirmed that there was a significant positive relation between self-explanation quality and the number of texts completed $F(1, 39) = 106.05$, $p < .001$, $R^2 = .731$. Specifically, Figure 5 illustrates that students improved their self-explanation quality as they interacted with a larger number of texts. In addition, those students with initially low performance improved such that they were indistinguishable from the initially high performing students. Students who performed poorly on the self-explanation pretest, compared to students who performed well on the self-explanation pretest, produced significantly lower quality self-explanations on the first 10 texts, but not after having received sufficient training (Jackson, Boonthum, & McNamara, 2010).

Insert Figure 5

This research and prior research with iSTART confirms that it effectively contributes to helping students improve their ability to self-explain and ultimately to better understand challenging science texts. In this regard, the roles of the NLP algorithms are key. First, they provide an integral part of the intelligence underlying the system. They drive the feedback which allows the system to respond to the student in intelligent adaptive ways. Second, the NLP algorithms promote engagement through verbal responses generated by the student. During the short-term interactions in the laboratory that lasted for two to four hours, these interactions were sufficient to engage and motivate the student. However, skill mastery requires long-term interaction with repeated practice (Newell & Rosenbloom, 1981). One side observation of this study in the classroom and one unfortunate side effect of long-term practice is that students become disengaged and uninterested in using the system when they use it over the course of weeks and months rather than hours (e.g., Bell & McNamara, 2007). Hence, more than just NLP algorithms were needed to motivate the students. For that reason, we turned to games. We developed iSTART-ME (Motivationally Enhanced) on top of the existing ITS by incorporating serious games and other game-based elements (Jackson, Boonthum, & McNamara, 2009; Jackson, Dempsey, & McNamara, 2010). In building this enhanced system, our hopes were to better motivate and engage the students such that they would be more apt to persist in iSTART in real world settings: in the classroom or at home.

iSTART-ME

The iSTART-ME game-based environment builds upon the existing iSTART system. The main goal of the iSTART-ME project is to integrate several game-based principles and features that are expected to support effective learning, increase motivation, and sustain engagement throughout a long-term interaction with an established ITS. The iSTART-ME system, along with theoretical justification for system design, has been extensively described in previous work (Jackson, Boonthum, & McNamara, 2009; Jackson, Dempsey, & McNamara, 2010), therefore only a brief description will be presented here.

The previous version of iSTART automatically progressed students from one text to another with no intervening actions. The new version of iSTART-ME is controlled through a selection menu (see Figure 6 for screenshot of the selection menu). Researchers claim that motivation and learning can be increased through multiple elements of a task including feedback, fantasy, personalization, choice, and curiosity (Cordova & Lepper, 1996; Papastergiou, 2009). Therefore, these features have been incorporated into the design of the iSTART-ME selection menu. This selection menu provides students with opportunities to interact with new texts, earn points, advance through levels, purchase rewards, personalize a character, and play educational mini-games (designed to use the same strategies as in practice).

Insert Figure 6

Several educational mini-games have been incorporated within iSTART-ME. In general, each of these mini-games has been designed so that a single session should be playable to completion within 10-20 minutes. The compilation of mini-games model strategy use and aim to improve: identification of strategies, generation of new self-explanations, meta-comprehension awareness, and/or vocabulary. Each mini-game focuses on one or two of these areas of improvement, and situates it within a game-based environment. After completion of a mini-game, students are directed back to the main iSTART-ME selection screen (see Figure 6).

Included in the selection menu, students can choose between three methods of generative practice (see Figure 7 for screenshots of Coached Practice, Showdown, and Map Conquest). All three methods utilize the previously described iSTART assessment algorithm and its corresponding output. Coached Practice is the updated version of the original iSTART practice, in which students are asked to generate their own self-explanations when presented with a text and specified target sentences. Students are guided through practice by Merlin, a wizard who reads sentences aloud, asks for a self-explanation at each target sentence, and provides verbal qualitative feedback for user-generated self-explanations. In addition, the new version of Coached Practice integrates basic game-based features, such as points and a feedback bar. For each submitted self-explanation, points are calculated based on both the current iSTART assessment score as well as the assessment score from the previous sentence. This scoring system was designed to reward consistent and quality performance, such that the maximum points are achieved by writing high quality self-explanations on consecutive sentences. Students are allowed to resubmit self-explanations after receiving feedback from Merlin. When multiple submissions are generated for a given target sentence, the average score across all submissions is used to determine the final point value for that sentence. The feedback bar provides students with a visual indication of the iSTART assessment score (i.e., 0=poor, 1=fair, 2=good, 3=great).

In Showdown, students compete against a computer player to win rounds by writing better self-explanations. After the student submits a self-explanation, it is scored, the quality assessment (iSTART self-explanation score) is represented as a number of stars (0-3 stars), and an opponent's self-explanation is also presented and scored. The self-explanation scores are compared and the player with the most stars wins the round. The player who wins the most rounds at the end of the game is declared the winner. Map Conquest is the other game-based method of practice where students generate their own self-explanations. In this game, the quality of a student's self-explanation determines the number of dice that student earns (0-3 dice). Students place these dice on a map, and use them to conquer neighboring opponent territories, which are controlled by two virtual opponents. It is worth noting that, unlike Coached Practice, students in Showdown and Map Conquest only write one self-explanation per target sentence and are not provided the opportunity to re-submit their self-explanation for a better score.

Insert Figure 7

There have been previous studies with the iSTART-ME game components that focused on single session studies and investigated individual elements within the system (Brunelle, Jackson, Dempsey, Boonthum, Levinstein, & McNamara, 2010; Dempsey, Jackson, Brunelle, Rowe, & McNamara, 2010). A more recent pilot study includes fewer participants who interacted with the full iSTART-ME system across multiple sessions spanning several weeks.

This study was designed to improve ecological validity and allow for student interactions that mimic how iSTART-ME could be implemented within a classroom environment. All participants (n=9) completed the full iSTART-ME training, including Introduction, Demonstration, Practice, and an extended interaction with the Selection Menu. Interactions with the system took place across eight different sessions (about an hour each) spanning three and a half weeks. After completing the initial training and Practice module, students spent the remainder of the sessions freely using all features within the Selection Menu. After interacting with iSTART-ME for 8 sessions, participants completed a posttest survey, which included questions about attitudes, enjoyment, and motivation. Figure 8 displays the average question ratings for the three generation environments (Jackson, Davis, Graesser, & McNamara, 2011).

Insert Figure 8

Coached Practice was consistently rated lower than one or both of the game-based practice methods. One of the most interesting results from these comparisons is the seemingly conflicting ratings for Map Conquest. This game was rated as significantly more frustrating than the other generation games; however, it was also rated as the most enjoyed generation game. Notably, whereas the participants reported that the map portion of the game was initially confusing (and therefore frustrating), it was also one of the most game-like and enjoyable aspects of the environment. A correlation between these ratings found that participants' frustration with the interface was negatively related to their enjoyment of Map Conquest, r=-.735, *p*=.024. Thus, updates that improve instructions (and avoid frustration), should yield even higher enjoyment ratings for this practice environment.

Discussion

The results from the current work are encouraging because they indicate a successful merging of two commonly problematic areas of educational research. This work focuses on creating an accurate assessment of performance during learning, and improving students' enjoyment and motivation during the learning process. The results support the current design of the iSTART-ME NLP algorithm, and indicate that students enjoyed interactions with the new game-based aspects of the system over an extended period of time. Specifically, the algorithm performance is comparable to human assessments, and allows the system to provide accurate and appropriate feedback. This combination has led to increased student performance with extended use of the system. Students' higher ratings for the game-based practice methods indicate that the new game additions to iSTART-ME improve enjoyment and will hopefully contribute to increased persistence over extended interactions. Indeed, these results are supported by additional longitudinal data comparing iSTART-ME (as described here) with the original non-game version of iSTART (Jackson & McNamara, in press).

One aspect of the mini-games provided in iSTART-ME is that they are fairly *primitive* in terms of aesthetics, particularly in comparison to today's gaming standards. Indeed, the screen shots provided in this paper may not induce the perception that these games would be particularly exciting. Certainly, an interesting question would be to compare these games to games (that would provide the same instruction) that are more sophisticated in terms of game technologies. Unfortunately, successful popular games are extremely costly and not within the budget of most funding agencies. Nonetheless, our research has indicated that students do enjoy the iSTART-ME games (see e.g., Jackson, Davis, Graesser, & McNamara, 2011; Jackson, Dempsey, & McNamara; 2012; Jackson & McNamara, in press), probably because they understand that the purpose of the games is not to have fun per se, but rather to have a more enjoyable experience while learning comprehension strategies.

Within the game-based analyses, one particularly interesting finding was that Map Conquest received the highest ratings for both frustration as well as enjoyment. Although the instructions and interface complexity of Map Conquest may have contributed to frustration (and lower ratings from some students), the majority of students persisted and provided high ratings of enjoyment for the game. These mixed ratings for Map Conquest further support the overall design of the iSTART-ME selection menu, which allows students to choose between a variety of games.

iSTART-ME can accurately assess student performance as well as successfully sustain user enjoyment over an extended amount of time. This finding provides a foundation for future work that more fully investigates the intricacies of assessment and the timelines of effects for specific game elements (e.g., competition, challenge, variety, control, etc.). Importantly, there are no other existing tutoring technologies that incorporate NLP within multiple games (or mini-games) and also have both game and non-game versions of practice. Because the iSTART-ME games vary in terms of game features and because the system itself is modular in nature, this puts it in a unique position to examine the benefits and detriments of game-based practice, as well as the differential effects of a variety of game elements. Our future work will focus on these issues as well as how these factors differentially affect both motivation and learning.

As shown in Figure 9, this work is conceptually driven by the assumption that game elements vary in terms of their potential effects on motivation and learning (McNamara, Jackson, & Graesser, 2010). Research has demonstrated that various mechanisms common to games, such as feedback, incentives, task difficulty, and control, can have a significant impact on motivation, 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; Schute, 2008). In turn, motivation is a multidimensional construct that subsumes a number of component factors, such as interest, enjoyment, expectancies, and values. *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 both intrinsic and extrinsic factors related to interest, engagement, enjoyment, self-regulation, and self-efficacy, which have been shown to positively impact learning (Alexander, Murphy, Woods, Duhon, & Parker, 1997; Bandura, 2000, Pajares, 1996; Pintrich, 2000; Zimmerman & Schunk, 2001).

Insert Figure 9

Based on the collective findings linking motivation and learning, Figure 9 provides a non-exhaustive visual mapping of empirically supported links, extending from sample *game features* through *interaction mechanisms* to *motivational constructs*, which in turn influence *behaviors and mental states* that support *learning and mastery* (individual relations are discussed in more detail within McNamara, Jackson, & Graesser, 2010). For example, across the top of Figure 1, the values associated with *points* and *levels* provide a user with *feedback* on their performance and progress through a system. Continuing along within the figure, there is abundant research in the cognitive area that has shown that various dimensions of *feedback* (structure, content, schedule, and delivery method) have a profound impact on the *learning* process and can influence both students' *self-efficacy* and *self-regulation* (Anderson et al., 1995; Corbett & Anderson, 1990; Foltz et al., 2000; Jackson & Graesser, 2007; Schunk & Pajares, 2001; Schute, 2008). Other examples from research have shown that incentives increase enjoyment (Moreno & Mayer, 2005), changing task difficulty can affect self-regulation (Boekarts & Cascallar, 2006; Schunk & Pajares, 2009), and providing control can improve interest (Cordova & Lepper, 1996). Overall, the concepts represented within Figure 9 have been examined within prior research with evidence suggesting that they should support an enjoyable and productive learning environment that sustains students' interest (Young et al., 2012).

The unique combination of work discussed here (NLP, ITS, and games) is the first step in a rapidly growing area of interdisciplinary research that can contribute to multiple research and educational communities. Allowing students to express themselves in natural language, combined with the added enjoyment from a game-based environment has the potential to greatly increase skill acquisition through a higher likelihood of interested, returning users (Garris, Ahlers, & Driskell, 2002; Gee, 2003; Steinkuehler, 2006).

References

Bell, C., & McNamara, D.S. (2007). Integrating iSTART into a high school curriculum. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 809-814). Austin, TX: Cognitive Science Society.Bransford, J., Brown, A., & Cocking, R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school.* Washington, D.C.: National Academy Press. Online at: http://www.nap.edu/html/howpeople1/

Brunelle, J.F., Jackson, G.T., Dempsey, K., Boonthum, C., Levenstein, I.B., & McNamara, D.S. (2010). Game-based iSTART practice: From MiBoard to self-explanation showdown. In H.W. Guesgen & C. Murray (Eds.), *Proceedings of the 23rd International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 480-485). Menlo Park, CA: The AAAI Press.

Cordova, D.I., & Lepper, M.R. (1996). Intrinsic motivation and the process of learning beneficial effects of contextualization, personalization and choice. *Journal of Educational Psychology, 88,* 715-730.

Dempsey, K., Jackson, G.T., Brunelle, J.F., Rowe, M.P., & McNamara, D.S. (2010). MiBoard: A digital game from a physical world. In H.W. Guesgen & C. Murray (Eds.), *Proceedings of the 23rd International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 498-503). Menlo Park, CA: The AAAI Press.

Garris, R., Ahlers, R., Driskell, J.E. (2002). Games, motivation and learning: A research and practice model. *Simulation and Gaming, 33*, 441-467.

Gee, J.P. (2003). What video games have to teach us about learning and literacy. New York: Palgrave Macmillian.

Graesser, A. C., Hu, X., & Person, N. (2001). Teaching with the help of talking heads. In T. Okamoto, R. Hartley, Kinshuk, J. P. Klus (Eds.), *Proceedings IEEE International Conference on Advanced Learning Technology: Issues, Achievements and Challenges* (pp. 460-461).

Jackson, G.T., Boonthum, C., & McNamara, D.S. (2009). iSTART-ME: Situating extended learning within a game-based environment. In H.C. Lane, A. Ogan, & V. Shute (Eds.), *Proceedings of the Workshop on Intelligent Educational Games at the 14th Annual Conference on Artificial Intelligence in Education* (pp. 59-68). Brighton, UK: AIED.

Jackson, G.T., Boonthum, C., & McNamara, D.S. (2010). The efficacy of iSTART extended practice: Low ability students catch up. In J. Kay & V. Aleven (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 349-351). Berling/Heidelberg: Springer.

Jackson, G.T., Davis, N.L., Graesser, A.C., & McNamara, D.S. (2011). Students' enjoyment of a game-based tutoring system. G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), the *Proceedings of the Artificial Intelligence in Education Society Conference* (pp. 475-477). Auckland: NZ: AIED.

Jackson, G.T., Dempsey K.B., & McNamara, D.S. (2010). The evolution of an automated reading strategy tutor: From classroom to a game-enhanced automated system. In M.S. Khine & I.M. Saleh (Eds.), *New Science of learning: Cognition, computers and collaboration in education* (pp. 283-306). New York, NY:Springer.

Jackson, G.T., Guess, R.H., & McNamara, D.S. (2010). Assessing cognitively complex strategy use in an untrained domain. *Topics in Cognitive Science*, *2,* 127-137.

Jackson, G. T., & McNamara, D. S. (in press). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology*.

Landauer, T., McNamara, D.S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.

Magliano, J.P., Todaro, S., Millis, K.K., Wiemer-Hastings, K., Kim, H.J., & McNamara, D.S. (2005). Changes in reading strategies as a function of reading training: A comparison of live and computerized training. *Journal of Educational Computing Research, 32*, 185–208.

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

McNamara, D.S., Boonthum, C., Levinstein, I.B., & Millis, K.K. (2007). Evaluating self-explanations in iSTART: comparing word-based and LSA algorithms. In T. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 227-241). Mahwah, NJ: Erlbaum.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers, 36*, 222–233.

McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research, 34,* 147–171.

McNamara, D.S., & Scott, J.L. (1999). Training reading strategies. In M. Hahn & S.C. Stoness (Eds.), *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society* (pp. 387-392). Hillsdale, NJ: Erlbaum.

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ.

O'Reilly, T., Best, R., & McNamara, D.S. (2004). Self-explanation reading training: Effects for low-knowledge readers. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Cognitive Science Society* (pp. 1053-1058). Mahwah, NJ: Erlbaum.

O'Reilly, T.P., Sinclair, G.P., & McNamara, D.S. (2004). Reading strategy training: Automated versus live. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the 26th Annual Cognitive Science Society* (pp. 1059-1064). Mahwah, NJ: Erlbaum.

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers and Education, 52*, 1-12.

Renner, A., McCarthy, P. M., Boonthum-Denecke, C., & McNamara, D. S. (2011). Maximizing ANLP evaluation: Harmonizing flawed input. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 438-456). Hershey, PA: IGI Global.

Steinkuehler, C.A. (2006). Massively multiplayer online video gaming as participation in a discourse. *Mind Culture & Activity, 13*, 38-52.

Table 1. Examples of Self-Explanation Categories for the Target Sentence “Energy-storing molecules are produced on the inner folds.”

iSTART Score	Category	Example Self-Explanation	Example Tutor Feedback
0	Irrelevant	“Hello, I am a taco. This sentence is very boring, and the little wizard guy talks funny.”	“Let’s see if you can add more information that relates to the paragraph.”
1	Sentence-based	“The molecules holding on to the energy are created on the inner folds.”	“O.K. If you add a little more next time, it will be even better.”
2	Text-based	“These sentences say that the mitochondria’s inner membrane produces energy storing molecules.”	“That's pretty good.”
3	Global-based	“The inner folds develop energy-storing molecules that help store more energy for the plant and help it grow, survive, and reproduce.”	“I’m impressed!”

Figure Captions

Figure 1. NLP cycle of self-explanation practice and feedback in iSTART.

Figure 2. Screenshot of iSTART Coached Practice.

Figure 3. Correspondence between human evaluations of the self-explanations for 2 trained texts and the iSTART assessment algorithm.

Figure 4. Correspondence between human evaluations of the self-explanations for untrained texts and the iSTART assessment algorithm.

Figure 5. Average self-explanation scores as a function of prior ability and the number of texts explained.

Figure 6. Screenshot of iSTART-ME selection menu.

Figure 7. Screenshots of generation practice environments.

Figure 8. Mean ratings for post-survey questions for 3 generation games.

Figure 9. Mapping between features, mechanisms, constructs, behaviors, and learning.

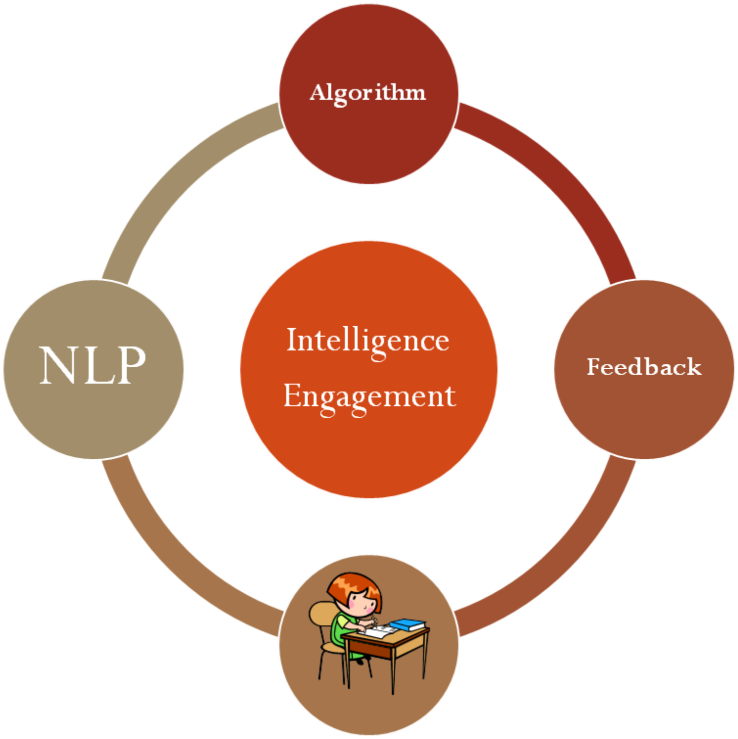
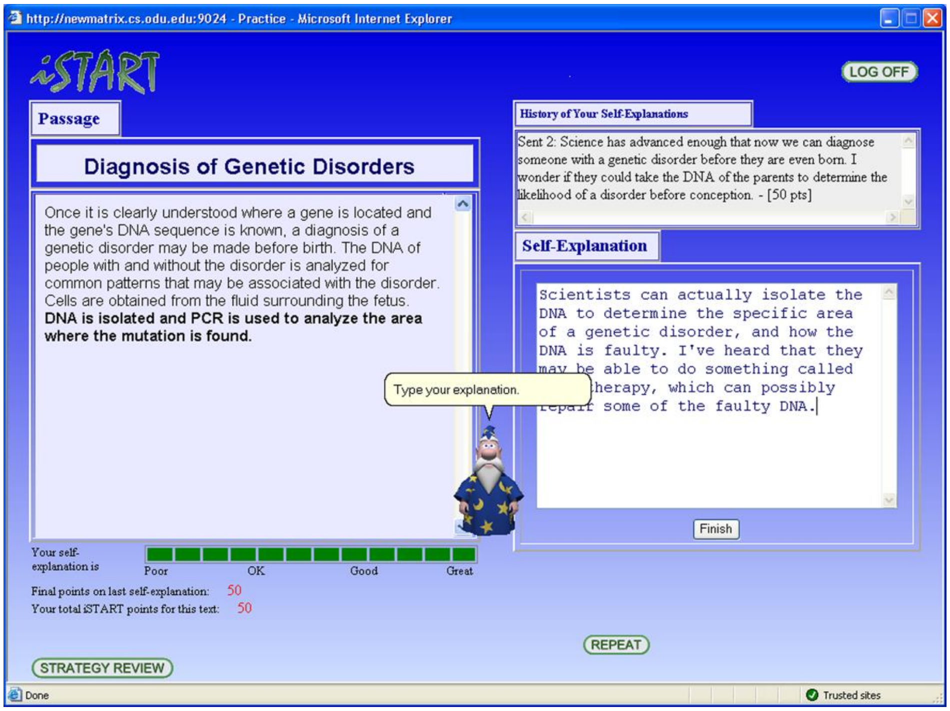


Figure 1. NLP cycle of self-explanation practice and feedback in iSTART.



Diagnosis of Genetic Disorders

Once it is clearly understood where a gene is located and the gene's DNA sequence is known, a diagnosis of a genetic disorder may be made before birth. The DNA of people with and without the disorder is analyzed for common patterns that may be associated with the disorder. Cells are obtained from the fluid surrounding the fetus. **DNA is isolated and PCR is used to analyze the area where the mutation is found.**

Sent 2: Science has advanced enough that now we can diagnose someone with a genetic disorder before they are even born. I wonder if they could take the DNA of the parents to determine the likelihood of a disorder before conception. - [50 pts]

Scientists can actually isolate the DNA to determine the specific area of a genetic disorder, and how the DNA is faulty. I've heard that they may be able to do something called therapy, which can possibly repair some of the faulty DNA.

Your self-explanation is

Final points on last self-explanation: 50

Your total iSTART points for this text: 50

Figure 2. Screenshot of iSTART Coached Practice.

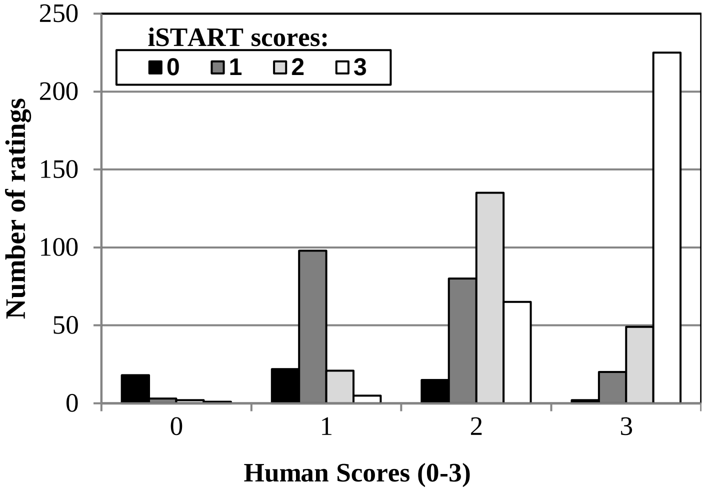


Figure 3. Correspondence between human evaluations of the self-explanations for 2 trained texts and the iSTART assessment algorithm.

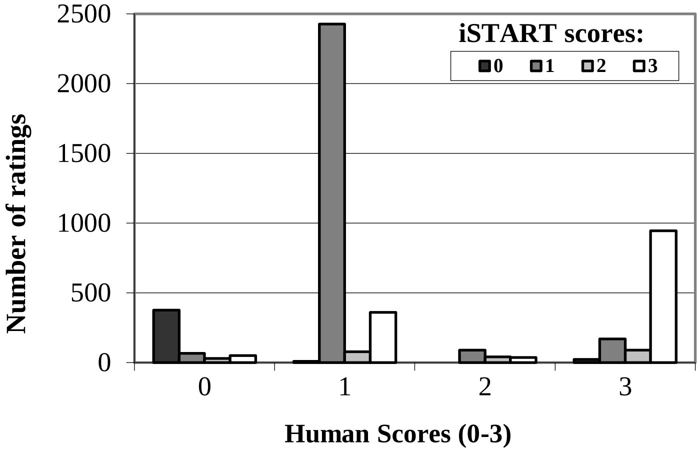


Figure 4. Correspondence between human evaluations of the self-explanations for untrained texts and the iSTART assessment algorithm.

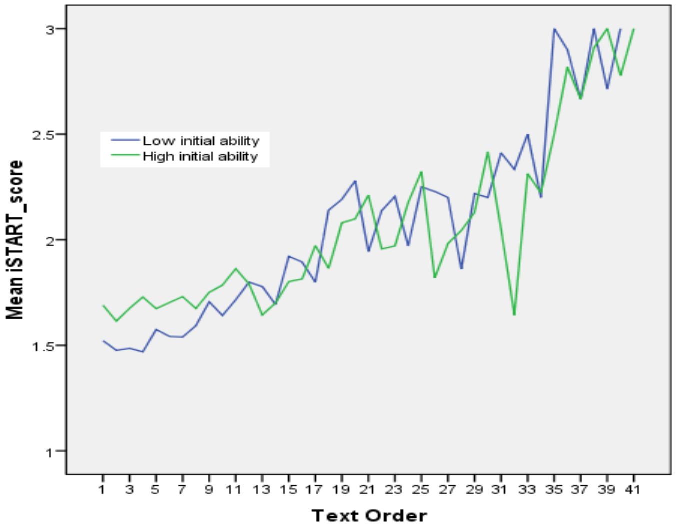
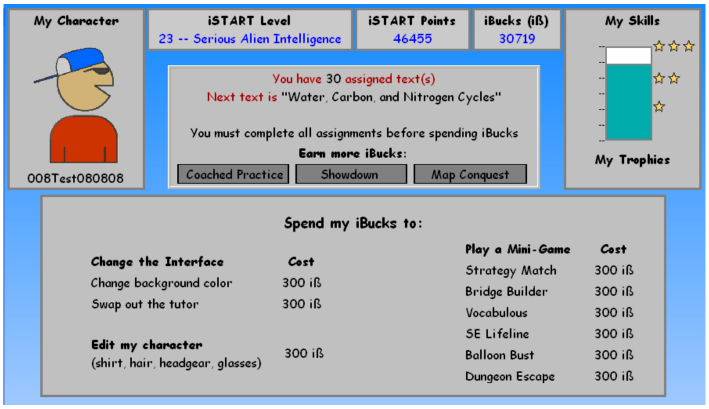


Figure 5. Average self-explanation scores as a function of prior ability and the number of texts explained



008Test080808

iSTART Level
23 -- Serious Alien Intelligence

iSTART Points
46455

iBucks (iß)
30719

You have 30 assigned text(s)
Next text is "Water, Carbon, and Nitrogen Cycles"

You must complete all assignments before spending iBucks

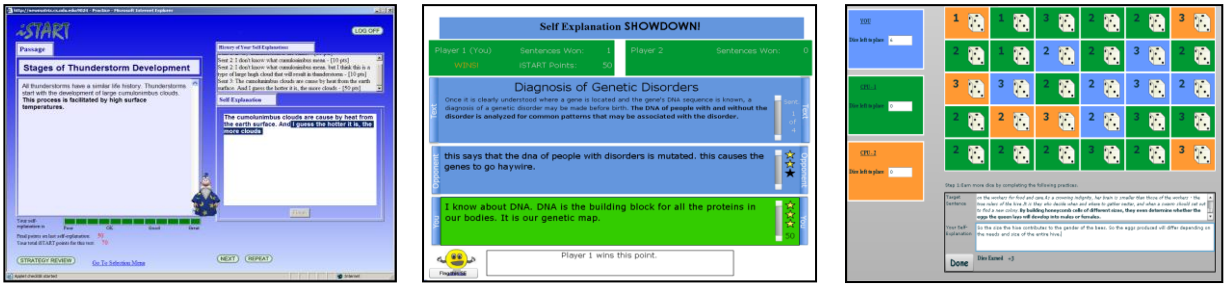
Earn more iBucks:

Spend my iBucks to:

Change the Interface	**Cost**
Change background color	300 iß
Swap out the tutor	300 iß
Edit my character (shirt, hair, headgear, glasses)	300 iß

Play a Mini-Game	**Cost**
Strategy Match	300 iß
Bridge Builder	300 iß
Vocabulous	300 iß
SE Lifeline	300 iß
Balloon Bust	300 iß
Dungeon Escape	300 iß

Figure 6. Screenshot of iSTART-ME selection menu.



Stages of Thunderstorm Development

All thunderstorms have a similar life history. Thunderstorms start with the development of large cumulonimbus clouds. **This process is facilitated by high surface temperatures.**

Sent 2: I don't know what cumulonimbus mean - [10 pts]
Sent 2: I don't know what cumulonimbus mean. but I think this is a type of large hugh cloud that will result in thunderstorms - [10 pts]
Sent 3: The cumolunimbus clouds are cause by heat from the earth surface. And I guess the hotter it is, the more clouds - [50 pts]

The cumolunimbus clouds are cause by heat from the earth surface. And I guess the hotter it is, the more clouds

Your self-explanation is:

Final points on last self-explanation: 50
Your total iSTART points for this text: 70

Player 1 (You) Sentences Won: 1
WINS! iSTART Points: 50
Player 2 Sentences Won: 0

Diagnosis of Genetic Disorders

Once it is clearly understood where a gene is located and the gene's DNA sequence is known, a diagnosis of a genetic disorder may be made before birth. **The DNA of people with and without the disorder is analyzed for common patterns that may be associated with the disorder.**

Sent. 1 of 4

this says that the dna of people with disorders is mutated. this causes the genes to go haywire.

I know about DNA. DNA is the building block for all the proteins in our bodies. It is our genetic map.

50

Player 1 wins this point.

YOU Dice left to place: 6
CPU - 1 Dice left to place: 0
CPU - 2 Dice left to place: 0

1	1	3	2	2	3
2	1	2	2	3	2
3	3	2	2	3	2
2	2	3	2	3	3
2	2	2	3	2	3

Step 1: Earn more dice by completing the following practices.

Target Sentence: on the workers for food and care. As a crowning indignity, her brain is smaller than those of the workers - the true rulers of the hive. It is they who decide when and where to gather nectar, and when a swarm should set out to find a new colony. **By building honeycomb cells of different sizes, they even determine whether the eggs the queen lays will develop into males or females.**

Your Self-Explanation: So the size the hive contributes to the gender of the bees. So the eggs produced will differ depending on the needs and size of the entire hive.

Dice Earned +3

Coached Practice Showdown Map Conquest

Figure 7. Screenshots of generation practice environments.

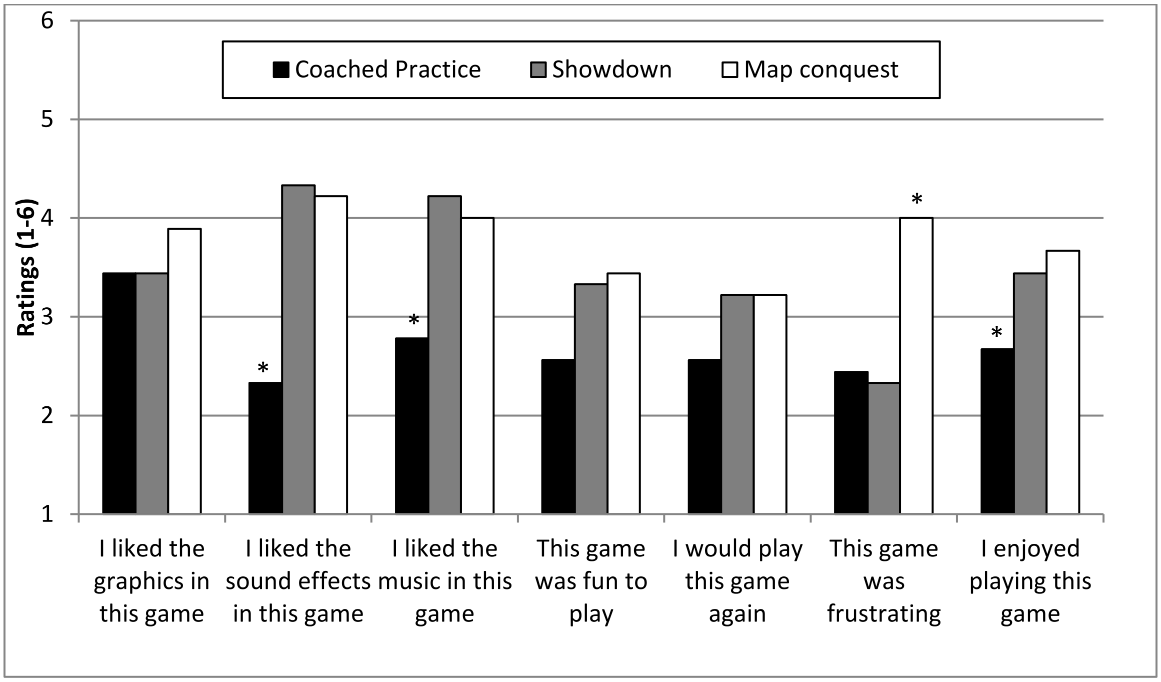


Figure 8. Mean ratings for post-survey questions for 3 generation games.

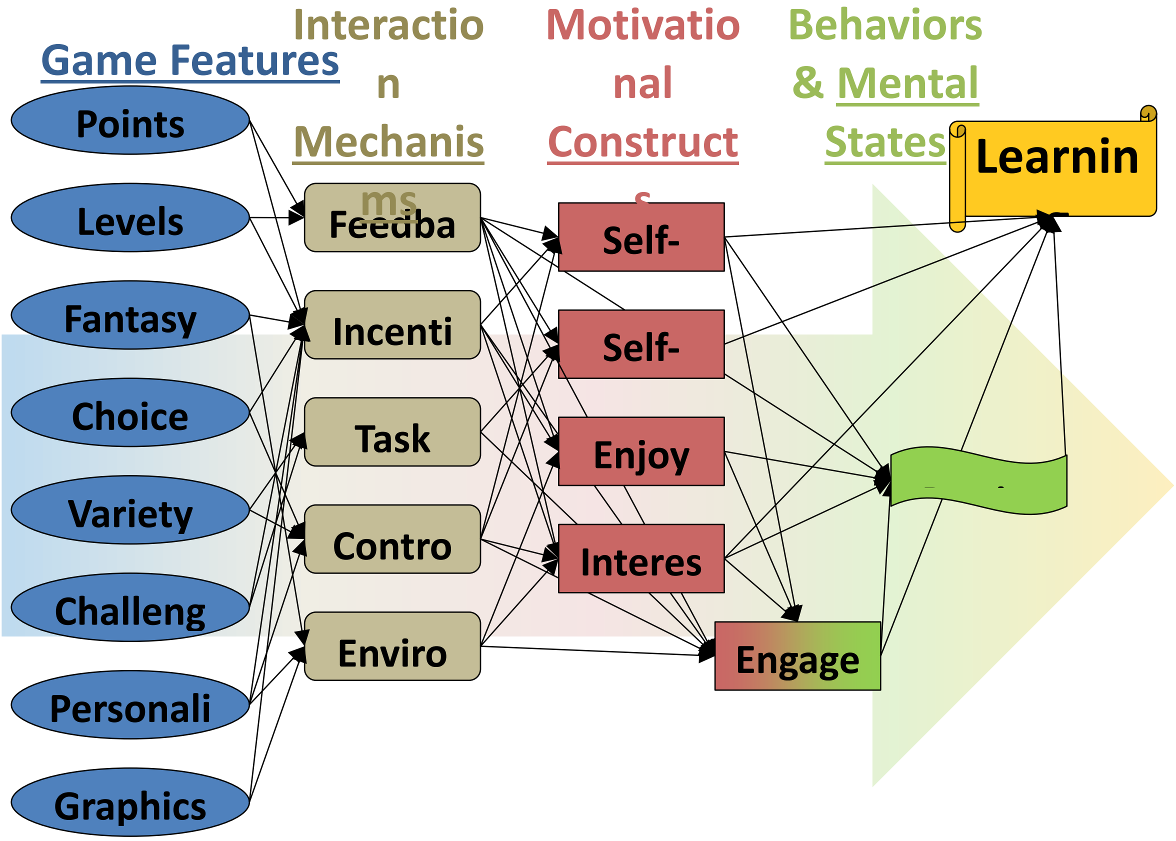


Figure 9. Mapping between features, mechanisms, constructs, behaviors, and learning.