Personalized Learning in iSTART: Past Modifications and Future Design

Kathryn S. McCarthy[1], Micah Watanabe[2], Jianmin Dai[2], & Danielle S. McNamara[2]

[1]Georgia State University, Department of Learning Sciences
[2]Arizona State University, Department of Psychology

To be submitted to the Personalization in Technology-Enhanced Learning Environments Special Issue of *Journal of Research on Technology in Education*

Accepted for publication: January 9th, 2020

Acknowledgements:

This project was funded by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A130124, R305A120707, R305A180261, R305A180144, and the Office of Naval Research, through Grant N00014140343, to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or the Office of Naval Research. The authors acknowledge the extensive contributions by Amy Johnson and Haojun Pei on this project, and also thank Matt Jacovina, Laura Allen, Renu Balyan, Kathleen Corley, Aaron Likens, Tricia Guerrero, Kevin Kent, Cecile Perret, Melissa Stone, Joseph Aubele, Carson Flood, Ashleigh Horowitz, Gary Ma, Amber Poteet, and our teacher-partners for their contributions to iSTART.

Corresponding Author:

Kathryn S. McCarthy
Department of Learning Sciences
Georgia State University
College of Education and Human Development
P.O. Box 3978
Atlanta, GA 30302
Kmccarthy12@gsu.edu

Abstract

Computer-based learning environments (CBLEs) provide unprecedented opportunities for personalized learning at scale. One such system, iSTART (Interactive Strategy Training for Active Reading and Thinking) is an adaptive, game-based tutoring system for reading comprehension. This paper describes how efforts to increase personalized learning have improved the system. It also provides results of a recent implementation of an adaptive logic that increases or decreases text difficulty based on students' performance rather than presenting texts randomly. High school students who received adaptive text selection showed increased sense of learning. Adaptive text selection also resulted in greater pre-training to post-training comprehension test gains, especially for less-skilled readers. The findings demonstrate that system-driven, just-in-time support consistent with the goals of personalized learning benefit the efficacy of computer-based learning environments.

Personalized Learning in iSTART: Past Modifications and Future Design

Recent advances in artificial intelligence and natural language processing have afforded a growing array of computer-based learning environments (CBLEs) aimed at improving literacy skills such as reading and writing (Crossley & McNamara, 2014; Passonneau et al., 2018). A key aspect of computer-based learning environments is that they can go beyond general instruction targeted toward the "average" student to provide rapid, individualized feedback that personalizes the students' learning experience.

Personalized Learning

While the idea of personalized learning is not new, the development of educational technologies has made it far easier to provide personalization in varied contexts and to diverse learners who differ across a wide array of dimensions, such as skills, knowledge, and motivation. The popularity of personalized learning has grown as policy-makers, researchers, and instructors acknowledge the need for adapting instruction to an increasingly diverse student populations. At the time of writing this manuscript, "personalized learning" yielded nearly 2.7 million hits in a Google search. Despite its popularity, there is ambiguity around what personalized learning *is* and how it affects student outcomes. Our own vision of personalized learning aligns most closely with that described by the U.S. Department of Education, which emphasizes that *personalized learning environments are those that are tailored to an individual learner's strengths, interests, and needs* (2010, 2016).

Critically, personalized learning is not merely adaptive to learners' skills. Personalized learning involves activities that self-initiated and are of personal relevance or interest (Walkington & Bernacki, 2018). Personalized learning is student-centered. Rather than the students being passive recipients of knowledge or skills, students shape their own learning based on their personal goals and interests. Personalized learning environments aim to facilitate and optimize conditions that afford and enhance student-centered learning. Relevant to the current work, adaptive learning platforms personalize learning by providing individualized feedback to students based on their knowledge or skill, generally in comparison to a theoretical model of student learning (e.g., Aleven & Koedinger, 2013; Koedinger & Corbett, 2006).

Personalized learning *can* include adaptive feedback. However, personalized learning environments consider other aspects of the student and context that afford tailoring the learning experience to the individual. As such, evaluations of CBLEs that leverage personalization must consider not only traditional learning outcomes, but the effects on aspects of self-regulated learning, such as motivation, engagement, and self-efficacy (Azevedo et al., 2009; D'Mello, Lehman, & Graesser, 2011; Walkington & Bernacki, 2014).

The current work describes how principles of personalized learning have been used to guide the iterative development of a computer-based, adaptive reading comprehension tutor. iSTART (Interactive Strategy Training for Active Reading and Thinking; McNamara, Levinstein, & Boonthum, 2004) provides adaptive, interactive, game-based tutoring to enhance adolescent students' comprehension of challenging content texts (e.g., science, history). We first describe iSTART and the need for such a system to improve students' comprehension skills. We then describe how principles of personalized learning have driven recent iterations of the system. Finally, we describe the design, implementation, and evaluation of a new system-driven personalization: adaptive text selection. We end with a discussion on personalization, and specific modifications to future versions of iSTART that will further increase the individualization of students' learning experience.

The Need for Reading Comprehension Strategy Instruction

Recent NAEP results indicate that more than half of American 12th graders struggle with reading comprehension (NAEP, 2015). Across a wide variety of grades and situations, strategy instruction has been shown to positively impact students' reading comprehension skill (Joseph et al., 2016; McNamara, 2011; Meyer & Ray, 2017; Pearson & Billman, 2016). Despite these findings and a growing emphasis on reading across the curriculum (Goldman et al., 2016; Horning, 2007), content instructors view reading instruction as outside of their duties and spend little time teaching students strategies that can help them to learn the content. For example, Ness (2016) found that secondary school content teachers spent less than 3% of instructional time on reading comprehension strategies.

There are two main reasons that students may not receive adequate comprehension strategy instruction. The first is that instructors often feel underprepared to teach literacy skills (Graham, Capizzi, Harris, Hebert, & Morphy, 2014) or feel that teaching comprehension is outside the scope of their duty as a content instructor (Ness, 2016). The second is that they simply do not have the time or resources to provide the kinds of iterative assessment and feedback necessary to cultivate these skills. This is especially true as students can vary widely on the type of support that they need. For example, students may vary in terms of their general reading skill, their prior knowledge of the domain, or their interest in the specific topic at hand. Instructors undoubtedly want to equip their students with the skills they need for success. However, asking them to provide personalized instruction for every student in their class in addition to teaching the required content is a tall order.

Thus, there is a gap that educational technologies for literacy are well-suited to fill. As described earlier, educational technologies such as CBLEs can provide effective instruction quickly and at scale. Automated tutors have been shown to be as effective as human tutors for individual students or small groups (Kulik & Fletcher, 2016; Ma, Adesope, Nesbit, & Liu, 2014; Steenbergen-Hu & Cooper, 2014; VanLehn, 2011). However, these studies generally describe CBLEs for well-defined domains, such as physics or math, in which questions can be scored as correct or incorrect and misconceptions can be quickly diagnosed with good foils. Far fewer adaptive learning technologies designed for ill-defined domains (Jacovina & McNamara, 2016; Jacovina, Snow, Dai, & McNamara, 2015; Lynch et al., 2006; Strobl et al., 2019).

In the past, the central obstacle for effective automated literacy instruction was the complexity of assessing open-ended responses (e.g., verbal protocols, essays) that do not have a singular “right” answer. Fortunately, the recent spike in the availability and sophistication of natural language processing (NLP) tools has allowed researchers to evaluate language efficiently and accurately. These NLP analyses can provide scores consistent with human raters (Jackson, Guess, & McNamara, 2010; Likens, Allen, McCarthy, & McNamara, 2018). The NLP data can also be used to model more latent aspects of students’ understanding (Allen, Snow, & McNamara, 2015), which can potentially drive relevant feedback that may not be obvious with the naked eye.

iSTART

iSTART is an automated tutor that began as an in-person intervention called Self-Explanation Reading Training (SERT; McNamara, 2004, 2017). SERT was motivated by research in cognitive psychology and discourse comprehension on the benefits of self-explanation, or explaining the text to one’s self during reading (Chi et al., 1994; see also Bisra et al., 2018) in combination with research in educational psychology showing the benefits of providing comprehension strategy instruction and practice (Palincsar & Brown, 1984; see McNamara 2007). Skilled readers tend to generate self-explanations on their own (Coté, Goldman, & Saul, 1998; Wolfe & Goldman, 2005). Prompting less-skilled readers to self-explain leads to more active engagement with the text and increased comprehension (Allen, McNamara, & McCrudden, 2014; Bisra et al., 2018; Chi et al., 1994). The effects of self-explanation are further enhanced when readers, particularly these less-skilled readers, are given instruction and practice on how to generate higher quality self-explanations (Magliano et al., 2005; McNamara et al, 2007). These higher quality self-explanations lead to a more robust mental model of the information in the text and, by extension, deeper comprehension (McNamara. 2004; McNamara, 2017).

iSTART was developed in response to the need for computer-based reading comprehension instruction. While SERT produced positive gains in students' reading comprehension skill, its reach was limited by the rate at which the instructor could provide feedback to each pair of students. The research team turned to developing NLP technologies to build an automated tutor that could quickly and accurately assess the quality of students' self-explanations as well as provide individualized feedback messages; thus, allowing self-explanation training to have a broader reach and bigger impact.

iSTART learners watch a series of short videos that introduce the process and benefits of self-explanation and the different strategies that can be used during self-explanation to improve comprehension (Figure 1a). These evidence-based reading comprehension strategies are comprehension monitoring, paraphrasing, predicting, and inferencing (bridging and elaboration). After the videos, learners are directed to a guided practice module called *Coached Practice*. Students practice using the strategies that they have learned by generating self-explanations at various target sentences (Figure 1b). The NLP algorithm uses both word-based indices and latent semantic analysis to provide a score from 0-3 (McNamara, Boonthum, Levinstein, & Millis, 2007). Scores of 0 and 1 indicate that the self-explanation is too short or quite similar to the target sentence, whereas scores of 2 and 3 indicate that the learner is generating inferences that connect information from the target sentence to other information in the text or to their prior knowledge (Jackson et al., 2010). Thus, these scores help to identify what comprehension strategy the student is using. Based on this score and additional performance measures considered within the algorithm, a pedagogical agent provides formative feedback to help students improve the quality of their self-explanation. For example, a learner whose self-explanation includes many of the same words as the sentence they are self-explaining might be reminded that “a good way to remember what you’ve read is to put the ideas of the text into your own words”. In contrast, a student who is already putting the ideas into their own words (i.e., paraphrasing) might be encouraged to generate an inference with the message “Great! Now you have the beginning of your explanation. Add to that by thinking about how the ideas in this sentence relate to ideas in the previous sentences of this text”. Students are then given the opportunity to revise their self-explanation.

Notably, the act of generating inferences is inherently aligned with principles of personalized learning. Elaborative inferences require students to draw upon their existing knowledge and experience and connect it to the information presented in the text. Unlike answering multiple-choice questions, open-ended self-explanation allows the student to make reading personally-relevant. Responses are not marked as right or wrong. Instead, the NLP algorithm is designed to honor students’ unique perspectives, while still extracting information about the students’ personal learning processes.

The Evolution of Personalization in iSTART

The initial version of iSTART focused on the use and efficacy of providing automated, adaptive feedback to students' self-explanations (McNamara, et al., 2004). Though these versions of iSTART had positive effects on self-explanation quality and deep comprehension of scientific texts, students tended to become disengaged before they had reached mastery (Bell & McNamara, 2007). As a result, iSTART evolved into iSTART-Motivationally Enhanced, or iSTART-ME (Jackson, Boonthum, & McNamara, 2009; Jackson & McNamara, 2013). Several aspects of the system were redesigned to improve motivation and engagement, in addition to the desired comprehension skill outcomes. There was a particular emphasis on increasing personalization through increased feedback and incentives combined with a focus on increasing learners' agency. Thus, increased personalized learning was realized by emphasizing the co-construction of the learning experience by both the learner and system. Learning activities are guided by real-time data-driven decisions, but students can exert ownership, or agency, over the learning environment.

In earlier versions of iSTART, students were ushered through the system on a single track. A limitation of this design is that students were not able to choose what to do next. To address this, iSTART-ME was modified so that students uniformly viewed the video lessons and a demonstration round of Coached Practice, but were then introduced to an open practice environment. In this environment, a student can continue to get feedback in additional rounds of Coached Practice or they can elect to play one of several self-explanation games (described below). They can also decide to go back to the video lessons for review or even view a progress page to see feedback about their system performance. The open practice environment allows each student to work on tasks of personal need or interest and still includes individualized feedback in each aspect of the environment.

The iSTART team investigated the effect of this open environment using random walk visualizations (Snow, Allen, Jacovina, & McNamara, 2015; Snow, Likens, Jackson, & McNamara, 2013). These walk diagrams showed that students took a variety of "paths" through iSTART-ME. Entropy analyses revealed that there were some path patterns that resulted in greater comprehension gains than others. What was most interesting, however, was that when students were forced into these "optimal" patterns of activity, they did not show comprehension gains demonstrated by those who selected the order to activities on their own. These results support the idea that the benefits of the open environment were related to students' sense of *agency* during learning.

The largest modification in the design of iSTART-ME was the introduction of a game-based practice environment (Jackson, Davis, Graesser, & McNamara, 2011; McNamara, Jackson, & Graesser, 2009; Figure 1c). iSTART-ME introduced generative games (students earn points by writing high quality self-explanations) and identification games (students earn points for identifying the correct strategy in an example; Figure 1d). High scores on these games are rewarded with trophies for exceptional scores and the ability to "level up" in the system. Leveling up unlocks new games. iSTART-ME also introduced a customizable player avatar and system environment that can be purchased points earned during gameplay. These "tangible" rewards increase motivation and allow for the student to co-create the learning environment.

Students responded positively to these system modifications. While both versions of iSTART were rated as equally helpful, iSTART-ME was rated as more motivating and enjoyable. Critically, as training went on, iSTART-ME was able to better sustain students' motivation (Jackson & McNamara, 2011, 2013). One concern was that these game elements would detract from meaningful strategy training. Indeed, the opportunity to engage with nongenerative practice (i.e., the identification games) and to modify aesthetic aspects of the system resulted in fewer self-explanations generated during practice. However, post-training self-explanation and comprehension test scores indicated no detrimental effect of reduced practice trials on performance (Jackson & McNamara, 2013). Ultimately, the introduction of the open, game-based environment had no direct effects on learning outcomes. However, it had positive effects on students motivation and enjoyment, which is a critical aspect of keeping students engaged in the long-term practice that they need to develop strong reading comprehension strategy skills.

Personalization through Text Selection Adaptivity

The modifications for iSTART-ME demonstrate the benefits of user-driven personalization. The purpose of the latest redesign, iSTART-3, was to focus on the addition of system-driven personalization that can adapt instruction and support to the student. When considering way to improve iSTART, it became clear that our adaptive and individualized instruction was still one-size-fits-all in a major aspect of system design: task difficulty. iSTART and iSTART-ME were programmed so that learners were presented texts in random order or an entire classroom would go through the texts in the order prescribed by the instructor.

One means of tailoring practice would be to simply provide texts that are at an appropriate grade level. There are, however, multiple problems with this approach. The first is that traditional metrics of readability (e.g., Flesch-Kincaid, Lexile) are generally poor indicators of comprehension performance (Duran et al., 2009). Indeed, Begeny and Green (2014) asked 360 students to read passages and answer questions from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) test and found that eight of the most common readability formulas were at or below chance at matching grade level with the students' actual performance. The second is that this approach assumes all students read at grade level and there is no variability in students' reading skill. Given that a majority of adult readers read at the 8th grade reading level, it is unlikely that "grade-appropriate" texts are actually appropriate for all of the high school and college students using iSTART.

A student reading well above grade level may find themselves bored, whereas a student reading below grade level may find it impossible to generate a high-quality self-explanation about the content. Thus, targeting tasks that match a students' current reading skill may have positive effects for both motivational and learning outcomes. Reading comprehension research indicates that matching texts to students' current needs benefits learning (Stanovich, 1986). However, given the limited use of CBLEs for ill-defined domains such as reading and writing, there is not a strong body of evidence to confirm the effects of automated text selection on learners' perceptions, behaviors, or learning outcomes.

Although the research in automated *text* adaptivity is limited, there is a large body of work on more general *task* adaptivity that selects the next task in instruction or practice based on some aspect of the student model such as their knowledge, affect, or effort (see Aleven, McLaughlin, Glenn, & Koedinger, 2016). Adapting the task has positive impacts as compared to random or uniform instruction (d = ~0.75; Van Lehn, 2011). Adapting the task based on readers' current knowledge or skill level affords the opportunity to target a students' "edge". Consistent with the cognitive notion of *desirable difficulties* (Bjork, 1994) as well as Vygotsky's *zone of proximal development* (ZPD; Vygotsky, 1978, 1981), tasks that are targeted toward learners' current abilities will be more effective than those that are too difficult or too easy. The desirable difficulties framework emphasizes that learning emerges from effortful processing. Consequently, tasks that are challenging are more beneficial for long-term effects than tasks that are easily accomplished (Bjork, 1994). Similarly, Vygotsky proposed the learning occurs most readily in the ZPD, the distance between what a learner can do independently and what they can do with additional support. The benefits of task adaptivity are not limited to simply overall learning gains. Adapting the task to the learner is also more efficient. Systems that adapt the task based on the current student model yield more learning in less time (Lovett, Meyer, & Thille, 2008; Stamper & Koedinger, 2011). Finally, and most critically for notions of personalized learning, adapting the task also has positive impacts on non-cognitive factors related to learning. CBLEs that target the ZPD have been shown to enhance student motivation (Arroyo, Woolf, Burelson, Muldner, Rai, & Tai, 2014; Brophy, 1999; Murray & Arroyo, 2002). Thus, exploring the benefits of adaptive text selection was a logical next step in our design modifications.

Given the differential effects of iSTART training for more and less skilled readers (e.g., Jackson, Boonthum, & McNamara, 2010; Magliano et al., 2005; McNamara et al., 2007), it was of interest to examine how text selection adaptivity might affect performance in the system. Adaptivity prevents skilled readers from losing interest in iSTART by providing sufficiently challenging tasks. Less skilled readers, benefit from adaptivity by working with texts that afford effortful and successful self-explanation attempts. Thus, adaptivity can have both a direct effect on the target skill as well as an indirect effect through increased self-efficacy. Higher self-efficacy, leads to prolonged motivation and persistence, which, in turn, yields increased opportunities for practice and feedback (e.g., Zimmerman, 2000).

With these benefits in mind, we designed and implemented an adaptive text selection algorithm for iSTART-3. We created a toggle that allowed us to turn on or turn off this algorithm. High school students were randomly assigned to iSTART training with adaptive text selection, iSTART training with texts selected randomly, or a no training (delayed treatment) control condition. We employed a multi-pronged approach to analysis, examining objective data via comprehension tests as well as self-reported user experience data, including overall enjoyment, frustration, and sense of learning and improvement. We also considered that personalized text selection might have varying impacts on different learners. Thus, we investigated general reading skill as a potential moderator.

We hypothesized that adaptive text selection would yield positive outcomes both in terms of students' experience with the system and in terms of enhanced reading comprehension gains.

H1. Adaptive text selection condition will increase engagement and motivation during iSTART training?

We predicted that personalization would be beneficial for students' non-cognitive outcomes based on findings that those working in their ZPD will have more sustained interest and effort (Arroyo et al., 2014).

H2a. Students in the adaptive text selection condition will show greater gains in self-explanation and comprehension as compared to who received practice and training with randomly selected texts.

We predicted that students who have sustained practice on tasks that are appropriately challenging (e.g., within their ZPD) will yield greater mastery of self-explanation skill (Ma et al., 2014; VanLehn, 2006; 2016). This will be reflected in improved self-explanation quality score and comprehension test scores from pretest to posttest.

H2b. The effect of adaptive text selection will be more pronounced for less-skilled readers.

Previous studies have demonstrated that the ways iSTART supports reading comprehension skill varies by reading skill (Magliano et al., 2005) and that system personalization can differentially impact learners of different reading skill (McCarthy et al., 2018). Although our sample is relatively small to detect moderating effects, it was important to examine the potential interactions between text selection adaptivity and reading skill. In particular, we hypothesized that the less-skilled readers would benefit from adaptive text selection more than skilled readers because adaptivity would afford them more time practicing and receiving feedback in their ZPD.

Method

Research Context

These data were collected in the context of a larger series of studies in the development of a new iSTART interface based on suggestions and input from our teacher-partners (iSTART-3; see McCarthy, Watanabe, & McNamara, under review). Consistent with the Design Implementation Framework (Stone et al., 2018), we conduct these developmental lab-based tests to empirically validate, and sometimes iteratively refine, design changes prior to full-scale ecological tests in classrooms. In the larger study, participants were randomly assigned to iSTART-3 training or a delayed treatment control. Given the focus of the present paper, we examine only the data from those in the training condition.

Participants & Design

The study was conducted in the spring and summer of 2018. Participants were 113 high school students from the southwestern United States. They were recruited through flyers and advertisements distributed in the area. Participants were given financial compensation for their participation. Participants were randomly assigned to one of two conditions: 1) iSTART training with random text selection (n = 32) or 2) iSTART training with adaptive text selection (n = 33). The sample (M_{age} = 16. 18, *SD* = 1.27; female = 51) self-identified as 38% Caucasian, 34%

Hispanic, 9% African-American, 8% Asian, and 11% identified as other. The majority (80%) were native English speakers.

Procedure

Participants completed the study over five sessions conducted in the research lab. In Session 1, all participants completed demographic survey and the individual difference measures, including Gates-MacGinitie Reading Test and the Learning Orientation, Performance Orientation Scale (see materials below). They then completed the self-explanation and comprehension pretest. Participants then completed three 2.5-hour sessions of iSTART training, including the lesson videos, Coached Practice, and practice games in which texts were either presented randomly or adaptively. At the end of each iSTART session, participants completed to a survey about their motivation and enjoyment.

Materials

iSTART Texts. To adapt text difficulty, we did not create more or less difficult versions of the same text. Instead, we assigned a difficulty level to each of 100 texts currently in the iSTART system library. These texts are aimed at middle, high school, and college students (see Jackson & McNamara, 2013). Traditional readability metrics (e.g., Flesch-Kincaid Grade Level) rely on superficial aspects of text, such as word and sentence length, that do not capture aspects of language that drive deep comprehension (Duran, Bellissens, Taylor, & McNamara, 2007; Graesser, McNamara, Louwerse, & Cai, 2004). To reflect more meaningful differences in text difficulty, we employed expert rankings to determine the degree to which any given text was more or less difficult than another. Through this process, the texts were ranked into nine levels roughly equivalent to grade level (6-14; for more details, see Balyan, McCarthy, & McNamara, 2018).

Text Selection Algorithm. The adaptive algorithm was designed based on students' average self-explanation score. In each generative game (Coached Practice, Map Conquest, Showdown), self-explanations are scored from 0-3. At the end of each game, if the learners' average self-explanation score was 2.0 or higher, the text difficulty level was raised one level. If the learners' average self-explanation score was less than 2.0, then the subsequent text was one level lower. Note that when a student plays an identification game, they stay at the same text difficulty level as they are not generating any self-explanations. A concrete example of the flow is shown in Figure 2. The student begins at the default start of level 10. On this text, they had an average self-explanation score of at least 2.0. Thus, the next text was a level 11 text. This ordered pattern of text difficult as compared to the data for a student who received randomly assigned texts shown in Figure 3.

While learners continually received feedback about their performance in terms of a self-explanation score, learners were not given any information about this text selection feedback loop. We implemented the adaptive text selection as a form "stealth scaffolding" because previous work has shown that making students overtly aware of their performance in iSTART had negative effects on their system performance and post-training outcomes (McCarthy, Likens, Johnson, Guerrero, & McNamara, 2018)[1].

Comprehension Tests. In the comprehension pretest and posttests, participants were prompted to self-explain while reading a scientific text, which was followed by a comprehension test. The texts and tests were adapted from those previously used in iSTART studies (e.g., Jackson & McNamara, 2011; McCarthy et al., 2018b; McNamara, O'Reilly, Best, & Ozuru, 2006). The texts, *Heart Disease, Red Blood Cells*, are approximately 300 words each and

[1] We did, however, build a stand-alone module, StairStepper, in which the difficulty of the text is not just overt, but central to the game mechanics (Perret, Johnson, McCarthy, Guerrero, & McNamara, 2017).

matched for linguistic difficulty. The comprehension tests were eight short-answer questions that probed for both memory for information in the text as well as readers' deeper understanding. The presentation of the texts as pretest or posttest were counterbalanced across participants.

Reading Skill Assessment. Participants completed the Gates-MacGinitie Reading Test (GMRT; MacGinitie & MacGinitie, 1989). The GMRT is a well-established measure of reading comprehension (α= .85-.92; Phillips, Norris, Osmond, & Maynard, 2002). The test is a 48-item multiple-choice test in which participants read short passages and answer two to six questions about each text.

Trait Motivation Scale. Trait level motivation was assessed using the Learning Orientation (LO) and Performance Orientation (PO) scales (Jha & Bhattachayya, 2013). Previous work has demonstrated that these factors have no effect on learning gains within previous version of iSTART (McCarthy, Likens, Kopp, Watanabe, Perret, & McNamara, 2018); nonetheless, a baseline measure of motivation was collected to contextualize differences in motivation that may manifest as a function of the text selection manipulation.

Daily Post-Training Motivation and Enjoyment Surveys. After each iSTART session, participants completed a seven-item survey adapted from (Jackson & McNamara, 2011). These items assess 1) students' overall experience, 2) negative experiences (boredom, frustration, and system problems), and 3) positive experiences (content learning, comprehension gains, enjoyment) during training (see Tables 1-3).

Results

We first report results from the daily experiences surveys to examine how adaptive text selection impacted student motivation and engagement. We then examine the pretest and posttest self-explanation and comprehension scores to evaluate the effect of text adaptivity on performance.

Motivation and Enjoyment

T-tests indicated no differences in trait motivation between conditions for LO score, $t(57) = .21$, $p = .83$, or PO score, $t(57) = .24$, $p = .81$.

Overall Experience. Participants reported that the sessions were generally fair to good [$M = 4.13$, $SD = 1.03$; using a Likert scale from (1) *very bad* to (6) *very good*; see Table 1]. To examine the effect of condition on participants' overall experience with iSTART, we conducted a 2(condition: random, adaptive) x 3(session: 1, 2, 3) repeated-measures ANOVA. This analysis revealed no change in reported experience from session to session, $F < 1.00$. The differences in response as a function of condition did not reach significance, $F(1, 56) = 1.70$, $p =.198$. The interaction between condition and session was not significant, $F < 1.00$.

Negative Experiences. To examine the effect of adaptivity on potential negative experiences (boredom, frustration, system errors), we conducted a series of 2(condition: random, adaptive) x 2(session: 1, 2, 3) ANOVAs. These analyses indicated no effects of session or condition, nor were there any significant interactions (see Table 2).

Boredom. Students reported that on average they were *sometimes* bored during the sessions ($M = 3.23$; $SD = .85$; using a Likert scale from (1) *never* to (5) *always*). The ANOVA indicated no main effect of session, $F <1.00$, nor condition $F(1,56) = 2.54$, p = .116. There was no interaction, $F(1, 112) = 1.57$, $p = .216$.

Frustration. Students reported that on average they were *rarely* to *sometimes* frustrated during the sessions ($M = 2.40$, $SD = .88$; using a Likert scale from (1) *never* to (5) *always*). The ANOVA revealed no significant effects, Fs < 1.00.

Problems. Students reported that on average they *rarely* experienced problems during the sessions (*M* = 2.05; *SD* = .80; using a Likert scale from (1) *never* to (5) *always*). The ANOVA indicated no main effect of session, *F* < 1.00. There was also no main effect of condition, *F*(1,56) = 1.77, *p* = .189, nor an interaction *F*(1,112) = 1.78, *p* = .188.

Positive Experiences. Another series of 2(condition: random, adaptive) x 2(session: 1, 2, 3) ANCOVAs, controlling for LO, were conducted to examine the effect of adaptivity on positive experiences with iSTART. There were no effects from session to session, nor interactions, but analyses indicated some perceived content learning benefits from the adaptive text selection condition (Table 3).

Learning. Students reported learning *a little* to *a lot* during the sessions (*M* = 3.31, *SD* = .72; using a Likert scale from (1) *not at all* to (5) *very much*; see Table 3). The ANOVA revealed a significant effect of session, such that students' reported learning decreased over the three sessions, *F*(1,56) = 13.87, *p* < .001. η_2 = .196. There was also a significant main effect of condition, *F*(1, 56) = 10.09, *p* = .002, η_2 = .15, such that those in the adaptive text selection condition reported learning more of the content (*M* = 3.59, *SD* = .63), than those who received texts randomly (*M* = 3.05, *SD* = .63). There was no interaction between session and condition, *F(*1,112) = 1.51, *p* = .224.

Reading Comprehension Improvement. Students reported improving in reading comprehension from *a little* to *a lot* during the sessions (*M*= 3.16; *SD* = .96; using a Likert scale from (1) *not at all* to (5) *very much*). There was a main effect of session, *F*(1,56) = 6.61, nor of condition, *F*(1,56) = 2.48, *p* = .121. There was also no interaction, *F*(1,112) = 3.03, *p* = .087.

Enjoyment. Students reported enjoyment of each session varied from *a little* to *a lot* (*M* = 3.08; *SD* = 1.09; using a Likert scale from (1) *not at all* to (5) *very much*). The ANOVA revealed no main effect of session, *F* < 1.00. There was also no main effect of condition, *F*(1,56) = 1.92, *p* = .171, nor an interaction *F*(1,112) = 1.62, *p* = .208.

Pretest and Posttest Outcomes

Scoring. The self-explanations generated during pretest and posttest were scored holistically from 0 to 3. Two raters achieved reliability on 20% of the set (Cohen's Kappa: Heart Disease = .57, Red Blood Cells = .65). The same raters scored the remainder of all of the responses and disagreements were settled by discussion.

Comprehension items from the pretest and posttest were scored using rubrics developed in previous studies. These rubrics award partial credit (0.0, 0.25, 0.50, 0.75, or 1.0) for incomplete answers (McNamara et al., 2006; Ozuru, Briner, Kurby, & McNamara, 2013). Two raters achieved good reliability on 20% of the set (Cohen's Kappa: Heart Disease = .86, Red Blood Cells = .74). The same two raters then scored the remainder of the responses and disagreements were settled by a third rater.

Preliminary Analysis. Table 4 shows average Gates-MacGinitie (GMRT) score, self-explanation scores, and comprehension test scores as a function of condition. A t-test indicated no differences in reading skill across the two conditions, *t*(64) = 1.04, *p* = .30).

Table 5 shows correlations between outcomes. Reading skill was moderately to strongly correlated with self-explanation and comprehension test scores.

Self-Explanation Scores. A series of linear mixed effects (LME) models were conducted to examine the effect of training condition on self-explanation score. This analysis was conducted using the lme4 package (version 1.1-15; Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2017). Participant was entered as a random factor in all models. Other variables were added as fixed factors. The baseline model (M0) included only reading skill as a continuous factor. Model 1 included the effect of test (pretest, posttest). Model 2 added the effect of training condition (random, adaptive) and Model 3 included the interaction terms. As shown in Table 6, Models 1 and 3 significantly improved model fit. Table 7 shows the model summary for Model 3. Reading skill accounted for the majority of variance in self-explanation score. Test was also a significant predictor of self-explanation score indicating that iSTART training increased self-explanation score from pretest to posttest. This was qualified by a significant three-way interaction.

Comprehension Test Scores. A parallel series of linear mixed effects models were conducted for comprehension test scores. Again, Models 1 (test) and 3 (interactions) improved model fit (Table 8). A summary of model 3 (Table 9) reveals, once again, that reading skill was a strong predictor of comprehension test score. As predicted, the only other significant factor was the three-way test by condition by reading skill interaction. Although reading skill was entered as a continuous variable, Figures 4 and 5 represent reading skill as “low” and “high” skill (median splits) to better visualize this complex interaction. Less skilled readers showed a benefit of text adaptivity, whereas skilled readers who received adaptive text selection showed a slight disadvantage as compared to those who received randomly assigned texts.

These findings are consistent with previous work that iSTART is most effective for less skilled readers. The current results extend these findings by demonstrating that additional personalization during training (e.g., text adaptivity) was adaptive text difficulty seems to be particularly beneficial for the less skilled readers, who are the students who are in greatest need of comprehension strategy support.

Discussion

In the current work, we described how personalized learning has been realized in the Interactive Strategy Training for Active Reading and Thinking (iSTART).

Summary

The iSTART team has implemented a variety of mechanisms that allow students to engage in personally-relevant learning activities including targeted feedback, gamification, and an open practice environment to enhance agency.

Our most recent addition to iSTART takes into consideration that learning is most likely to occur at the zone of proximal development. As such, the latest version of iSTART, iSTART-3, employs all of the previous personalization features of iSTART-ME with the addition of an adaptive text selection that scaffolds students' strategy use by selecting texts that are "just right" in terms of difficulty. The results indicate that adaptive text selection increased students' sense of learning and did not negatively impact their experience with the system. Examination of learning outcomes (self-explanation quality; comprehension test score) indicated that participants' improved in both outcomes from pretest to posttest. There were no main effects of the adaptivity condition. However, this was qualified by a complex interaction indicating that the adaptivity was beneficial for less-skilled readers. We observed little improvement in comprehension scores for skilled readers. It is of note that some participants quickly ascended to reading the most difficult texts in the library and continued to generate high quality self-explanations. Thus, one potential explanation for this reading skill by training interaction is that the more skilled readers were practicing below their ZPD and not being sufficiently challenged. While it might be possible to increase the breadth of texts available, previous work shows that skilled readers spontaneously generate self-explanations (e.g., Cote, Goldman, & Saul, 1998). That is – these highly-skilled readers may simply not need self-explanation training at all. As such, we are pleased to see that the adaptive text selection aided those who need it most.

Participants' self-reports regarding daily motivation and engagement at the end of each session indicated that students' perceptions of iSTART were generally positive. These ratings are similar to those in previous studies that demonstrated overall positive feelings towards the game-based iSTART system (e.g., Jackson, & McNamara, 2013). It is of interest that there were few perceived differences between the two versions of iSTART. The lack of differences in either positive or negative experiences likely indicates that students were unaware that text difficulty was being manipulated. However, one difference emerged in students' self-report data. Participants reported stronger feelings of learning in the adaptive text selection condition as compared to the random text selection condition, with a moderate effect size (Cohen's $d = .61$), which tended to increase from Session 1 (Cohen's $d = .35$) to Sessions 2 and 3 (Cohen's $d = .77$ and .70). Although students are often overconfident in their metacognitive judgments (Dunlosky & Metcalfe, 2008), the ability to successfully perform a task can increase the sense of *learning* and self-efficacy. Increased self-efficacy can lead to increased interest and persistence (see Zimmerman, 2000). These results collectively provide promising evidence that students' perceptions of learning are increased with adaptive text selection, and learning is positively impacted, particularly for less skilled readers.

Limitations

An important consideration is that neither version of the iSTART allowed the learner to select which text to read next. That is, this system-driven personalization may have impeded upon aspects of user-driven personalization that have been shown to enhance agency (Nagle, Novak, Wolfe, & Riener, 2014; Snow, Allen, Jacovina, & McNamara, 2015). As we continue to iteratively design and modify iSTART to improve the system, it would be of value to explore the effects of allowing students choose their texts, rather than having them randomly assigned or chosen by a teacher or experimenter. Such an approach might allow students to choose texts that better match their interests. However, informal pilot studies have indicated that the process of choosing a text is excessively time consuming; students spend more time choosing the texts than engaging in the practice activities. This tension is an important consideration researchers as designers continue to explore how different aspects of personalization in educational technology can impact learning.

These early findings suggest a modest benefit of adaptive text selection. However, the limited sample size prevents us from make strong claims. We intend to follow-up this study with a larger sample to further investigate the relation between adaptive text selection and individual differences. We can further explore reading skill as well as well as target other individual differences known to impact performance such as prior knowledge.

It will also be critical to explore how iSTART-3 training with adaptive text selection compares to in-person tailored instruction, both in terms of efficacy and feasibility of implementation. That is, adaptive text selection may yield only modest effects in comparison to a one-on-one tutoring. However, the ability to provide automated scaffolding may be the only thing feasible at scale. Indeed, there is great need for further work examining how personalized learning in the context of CBLEs influences learning across contexts.

Implications

The evolution of iSTART serves as a powerful example of how computer-based learning environments can benefit from the thoughtful implementation of elements that support personalized learning. The ability to provide accurate automated evaluation and feedback for open-ended domains such as reading comprehension is in and of itself a testament to the potential of computer-based learning. However, thinking of technology in education as merely a way to efficiently dole out instruction undermines its potential. Indeed, technology-driven instruction offers the means of giving each student a meaningful, one-on-one experience that can enrich their learning. These findings provide additional evidence that personalization is not one size fits all and that some features will be more or less effective for different learners. From a theoretical perspective, examining individual differences allows researchers to better identify how cognitive and noncognitive aspects of learners can interact. Such work can provide further nuance in the way that personalization features are deployed.

Work in iSTART demonstrates how principles of personalization can increase the quality and efficacy of learning. Further, the iterative changes that we have describe exemplify the fact that personalization is not a single thing. That is, a system is more than simply “personalized or not”. There are multiple *dimensions* along which a CBLEs can be modified (Walkington & Bernacki, 2014) and varying *degrees* of personalization that can address student’s skills, interests, and needs. CBLE designers can and should implement and evaluate a variety of means of increasing personalization, in part, to understand potential interactions, but also to deliver experiences that place each individual learner at the center of their own experience.

References

Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Are you reading my mind? Modeling students' reading comprehension skills with Natural Language Processing techniques. In J. Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, *24*(4), 387-426.

Aleven, V., & Koedinger, K. R. (2013). Knowledge component (KC) approaches to learner modeling. *Design Recommendations for Intelligent Tutoring Systems*, *1*, 165-182.

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. *Handbook of research on learning and instruction. Routledge*.

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, 24(4), 387-426.

Azevedo, R., Witherspoon, A., Graesser, A.C., McNamara, D.S., Chauncey, A., Siler, E., Cai, Z., Rus, V., & Lintean, M. (2009). MetaTutor: Analyzing self-regulated learning in a tutoring system for biology. 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. 635-637). Amsterdam, The Netherlands: IOS Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Balyan, R., McCarthy, K. S., & McNamara, D. S. (2018). Comparing machine learning classification approaches for predicting expository text difficulty. In *Proceedings of the 31st Annual Florida Artificial Intelligence Research Society (FLAIRS-31).* Melbourne, FL: AAAI Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

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.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Biancarosa, G., & Snow, C. E. (2004). Reading next: A vision for action and research in middle and high school literacy: A report from Carnegie Corporation of New York. Alliance for Excellent Education.

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing Self-Explanation: A Meta-Analysis. *Educational Psychology Review*, *30*(3), 703-725.

Brophy, J. (1999). Toward a model of the value aspects of motivation in education: Developing appreciation for particular learning domains and activities. *Educational Psychologist, 34*(2), 75-85.

Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, *18*(3), 439-477.

Crossley, S. A., & McNamara, D. S. (2014). Developing component scores from natural language processing tools to assess human ratings of essay quality. In W. Eberle & C. Boonthum-Denecke (Eds.), *Proceedings of the 27th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 381-386). Palo Alto, CA: AAAI Press.

D'Mello, S. K., Lehman, B., & Graesser, A. (2011). A motivationally supportive affect-sensitive autotutor. In *New perspectives on affect and learning technologies* (pp. 113-126). Springer, New York, NY.

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Sage Publications.

Duran, N., Bellissens, C., Taylor, R., & McNamara, D. (2007). Qualifying text difficulty with automated indices of cohesion and semantics. In D.S. McNamara and G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society(pp. 233-238).* Austin, TX: Cognitive Science Society.

Goldman, S. R., Snow, C., & Vaughn, S. (2016). Common themes in teaching reading for understanding: Lessons from three projects. *Journal of Adolescent & Adult Literacy,* 60(3), 255-264.

Graesser, A.C., McNamara, D.S., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*, 193-202.

Graham, S., Capizzi, A., Harris, K. R., Hebert, M., & Morphy, P. (2014). Teaching writing to middle school students: A national survey. *Reading and Writing*, 27(6), 1015-1042.

Horning, A. S. (2007). Reading across the curriculum as the key to student success. *Across the disciplines*, 4, 1-18.

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., Davis, N. L., Graesser, A. C., & McNamara, D. S. (2011). Students' enjoyment of a game-based tutoring system. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 475-477). Auckland, New Zealand, AIED.

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. (2011). Motivational impacts of a game-based intelligent tutoring system. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 519-524). Menlo Park, CA: AAAI Press.

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

Jacovina, M., & McNamara, D. S. (2016). Intelligent tutoring systems for literacy: Existing technologies and continuing challenges. In R. Atkinson (Ed.), Intelligent tutoring systems: Structure, applications and challenges. Hauppauge, NY: Nova Science Publishers Inc.

Jacovina, M. E., Snow, E. L., Dai, J., & McNamara, D. S. (2015). Authoring tools for ill-defined domains in intelligent tutoring systems: Flexibility and stealth assessment. *Design Recommendations for Intelligent Tutoring Systems, 3*, 109-121.

Jha, S., & Bhattacharyya, S. S. (2013). Learning orientation and performance orientation: Scale development and its relationship with performance. *Global Business Review, 14*(1), 43-54.

Joseph, L. M., Alber-Morgan, S., Cullen, J., & Rouse, C. (2016). The effects of self-questioning on reading comprehension: A literature review. *Reading & Writing Quarterly*, *32*(2), 152-173.

Koedinger, K. R., & Corbett, A. (2006). Cognitive Tutors: Technology Bringing Learning Science to the Classroom.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of Educational Research*, *86*(1), 42-78.

Li, H., & Baer, W. (2018). Scaffolding adult learners' reading strategies in the intelligent tutoring system. In K. Millis, D. Long, J. Magliano, K. Wiemer (Eds). *Deep Comprehension* (pp. 166-179). Routledge.

Likens, A. D., McCarthy, K. S., Allen, L. K., & McNamara, D. S. (2018). Recurrence Quantification Analysis as a method for studying text comprehension dynamics. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (LAK'18). Sydney, Australia.

Lovett, M., Meyer, O., & Thille, C. (2008). The Open Learning Initiative: Measuring the Effectiveness of the OLI Statistics Course in Accelerating Student Learning. *Journal of Interactive Media in Education*.

Lynch, C. F., Ashley, K. D., Aleven, V., & Pinkwart, N. (2006). Defining ill-defined domains; a literature survey. In *Intelligent Tutoring Systems (ITS 2006): Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*(4), 901.

Magliano, J.P., Todaro, S. Millis, 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.

MacGinitie, W. H., & MacGinitie, R. K. (1989). *Gates–MacGinitie Reading Tests (3rd Ed.)*. Itasca, IL: Riverside.

McCarthy, K. S., Likens, A. D., Kopp, K. J., Watanabe, M., Perret, C. A., & McNamara, D. S. (2018a). The "LO"-down on grit: Non-cognitive trait assessments fail to predict learning gains in iSTART and W-Pal. In *Companion Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK'18).* Sydney, Australia.

McCarthy, K. S., Likens, A. D., Johnson, A. M., Guerrero, T. A., & McNamara, D. S. (2018b). Metacognitive Overload!: Positive and Negative Effects of Metacognitive Prompts in an Intelligent Tutoring System. *International Journal of Artificial Intelligence in Education*, 1-19.

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

McNamara, D. S. (2011). Measuring deep, reflective comprehension and learning strategies: Challenges and successes. *Metacognition and Learning*, 3, 1-11.

McNamara, D. S. (2017). Self-Explanation and Reading Strategy Training (SERT) Improves low-knowledge students' science course performance. *Discourse Processes*, *54*(7), 479-492.

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

McNamara, D.S., Boonthum, C., Levinstein, I.B., & Millis, 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., Jackson, G.T., & Graesser, A.C. (2009). Intelligent tutoring and games (iTaG). 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. 1-10). Brighton, UK: AIED.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral 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.

Mesmer, H. A. E. (2008). Tools for matching readers to texts: Research-based practices. Guilford Press.

Meyer, B. J., & Ray, M. N. (2017). Structure strategy interventions: Increasing reading comprehension of expository text. *International Electronic Journal of Elementary Education,* 4(1), 127-152.

Murray, T., & Arroyo, I. (2002, June). Toward measuring and maintaining the zone of proximal development in adaptive instructional systems. In S.A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.) *International Conference on Intelligent Tutoring Systems* (pp. 749-758). Springer, Berlin, Heidelberg.

Nagle, A., Novak, D., Wolf, P., & Riener, R. (2014, May). The effect of different difficulty adaptation strategies on enjoyment and performance in a serious game for memory training. In *Proceedings of the 2014 IEEE 3rd International Conference on Serious Games and Applications for Health* (pp. 1-8).

NAEP (2015) U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Ness, M. K. (2016). Reading comprehension strategies in secondary content area classrooms: Teacher use of and attitudes towards reading comprehension instruction. *Reading Horizons, 49*(2), 5.

O'Connor, R. E., Bell, K. M., Harty, K. R., Larkin, L. K., Sackor, S. M., & Zigmond, N. (2002). Teaching reading to poor readers in the intermediate grades: A comparison of text difficulty. Journal of Educational Psychology, 94(3), 474.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing text comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology, 67*, 215-227.

Passonneau, R. J., McNamara, D., Muresan, S., & Perin, D. (2017). Preface: special issue on multidisciplinary approaches to AI and education for reading and writing. *International Journal of Artificial Intelligence in Education, 27(*4), 665-670.

Pearson, P. D., & Billman, A. K. (2016). Reading to Learn Science: A Right That Extends to Every Reader—Expert or Novice. In *Human Rights in Language and STEM Education* (pp. 17-34). Brill Sense.

Perret, C. A., Johnson, A. M., McCarthy, K. S., Guerrero, T. A., & McNamara, D. S. (2017). StairStepper: An adaptive remedial iSTART module. In B. Boulay, R. Baker & E. Andre (Eds.), *Proceedings of the 18th International Conference on Artificial Intelligence in Education (AIED),* (pp.557-560), Wuhan, China: Springer.

Phillips, L. M., Norris, S. P., Osmond, W. C., & Maynard, A. M. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. *Journal of Educational Psychology*, *94*(1), 3-13.

Snow, E. L., Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2015). Does agency matter?: Exploring the impact of controlled behaviors within a game-based environment. *Computers & Education*, *26*, 378-392.

Snow, E. L., Likens, A., Jackson, G. T., & McNamara, D. S. (2013). Students' walk through tutoring: Using a random walk analysis to profile students. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 276-279). Heidelberg, Berlin, Germany: Springer.

Stamper, J. C., & Koedinger, K. R. (2011, June). Human-machine student model discovery and improvement using DataShop. In *International Conference on Artificial Intelligence in Education* (pp. 353-360). Springer, Berlin, Heidelberg.

Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. Reading Research Quarterly, 2(4), 360-407.

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, 106(2), 331.

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education*, *131*, 33-48.

U.S. Department of Education (2010). *Transforming American Education: Learning Powered by Technology.* Office of Educational Technology, Washington, D.C. http://www.ed.gov/sites/default/files/netp2010.pdf

U.S. Department of Education (2016). *Future Ready Learning: Reimagining the Role of Technology in Education.* Office of Educational Technology, Washington, D.C. http://tech.ed.gov/files/2015/12/NETP16.pdf

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227-265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197-221.

VanLehn, K. (2016). Regulative loops, step loops and task loops. *International Journal of Artificial Intelligence in Education*, *26*(1), 107-112.

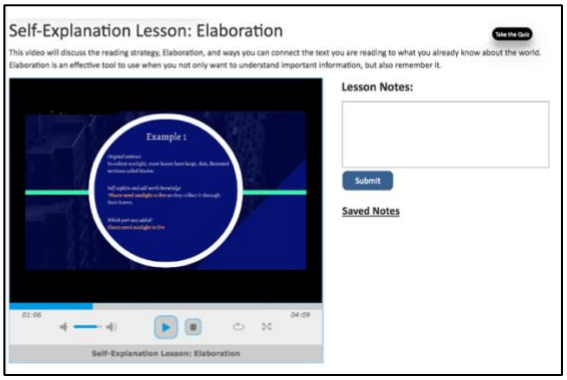
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1981). The instrumental method in psychology. In J. V Wertsch (Ed.), *The concept of activity in Soviet psychology* (pp. 134-144). Armonk, NY: Sharpe.

Walkington, C., & Bernacki, M. (2014). Motivating students by “personalizing” learning around individual interests: A consideration of theory, design, and implementation issues. In S. Karabenick & T. Urdan (eds.) *Advances in Motivation and Achievement Volume 18* (pp. 139-176), Emerald Group Publishing.

Walkington, C., & Bernacki, M. L. (2018). Personalization of instruction: Design dimensions and implications for cognition. *The Journal of Experimental Education*, *86*(1), 50-68.

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary educational psychology*, *25*(1), 82-91.



Self-Explanation Lesson: Elaboration

This video will discuss the reading strategy, Elaboration, and ways you can connect the text you are reading to what you already know about the world. Elaboration is an effective tool to use when you not only want to understand important information, but also remember it.

Example 1

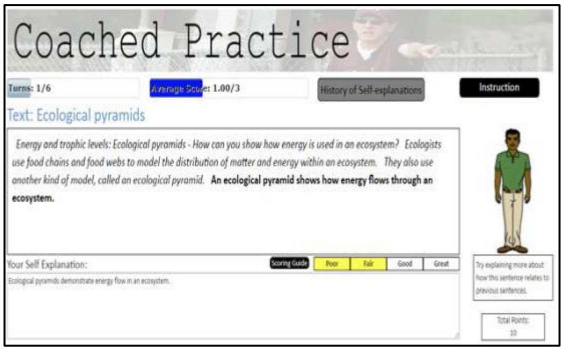
Original sentence
To collect sunlight, most leaves have large, thin, flattened sections called blades.

Self-explain and add world knowledge
"Plants need sunlight to live so they collect it through their leaves."

Which part was added?
Plants need sunlight to live

Lesson Notes:

Figure 1a. Video Lesson (Elaboration Strategy)



Coached Practice

Turns: 1/6

Average Score: 1.00/3

Text: Ecological pyramids

Energy and trophic levels: Ecological pyramids - How can you show how energy is used in an ecosystem? Ecologists use food chains and food webs to model the distribution of matter and energy within an ecosystem. They also use another kind of model, called an ecological pyramid. **An ecological pyramid shows how energy flows through an ecosystem.**

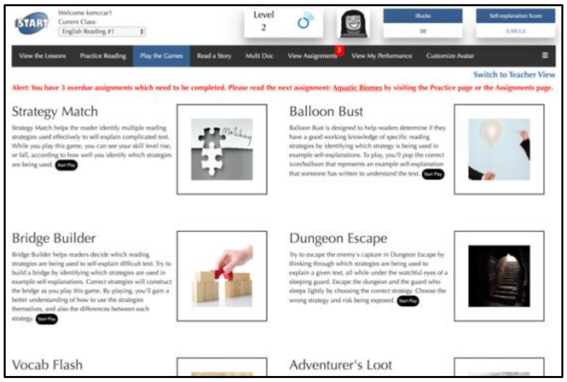
Your Self Explanation:

Ecological pyramids demonstrate energy flow in an ecosystem.

Try explaining more about how this sentence relates to previous sentences.

Total Points: 10

Figure 1b. Coached Practice



Welcome kmccar1
Current Class:
English Reading #1

Level 2

iBucks 10

Self-explanation Score

Alert: You have 3 overdue assignments which need to be completed. Please read the next assignment: Aquatic Biomes by visiting the Practice page or the Assignments page.

Strategy Match

Strategy Match helps the reader identify multiple reading strategies used effectively to self-explain complicated text. While you play this game, you can see your skill level rise, or fall, according to how well you identify which strategies are being used. Start Play

Balloon Bust

Balloon Bust is designed to help readers determine if they have a good working knowledge of specific reading strategies by identifying which strategy is being used in example self-explanations. To play, you'll pop the correct icon/balloon that represents an example self-explanation that someone has written to understand the text. Start Play

Bridge Builder

Bridge Builder helps readers decide which reading strategies are being used to self-explain difficult text. Try to build a bridge by identifying which strategies are used in example self-explanations. Correct strategies will construct the bridge as you play this game. By playing, you'll gain a better understanding of how to use the strategies themselves, and also the differences between each strategy. Start Play

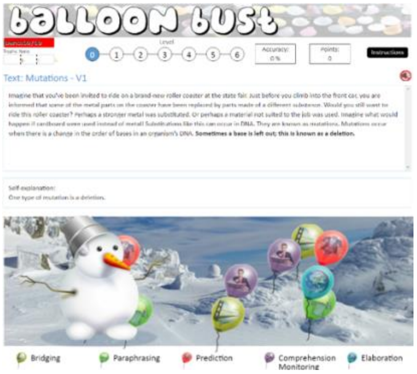
Dungeon Escape

Try to escape the enemy's capture in Dungeon Escape by thinking through which strategies are being used to explain a given text, all while under the watchful eyes of a sleeping guard. Escape the dungeon and the guard who sleeps lightly by choosing the correct strategy. Choose the wrong strategy and risk being exposed. Start Play

Vocab Flash

Adventurer's Loot

Figure 1c. Game Selection Interface



balloon bust

Text: Mutations - V1

Imagine that you've been invited to ride on a brand-new roller coaster at the state fair. Just before you climb into the front car, you are informed that some of the metal parts on the coaster have been replaced by parts made of a different substance. Would you still want to ride this roller coaster? Perhaps a stronger metal was substituted. Or perhaps a material not suited to the job was used. Imagine what would happen if cardboard were used instead of metal! Substitutions like this can occur in DNA. They are known as mutations. Mutations occur when there is a change in the order of bases in an organism's DNA. **Sometimes a base is left out; this is known as a deletion.**

Self-explanation:
One type of mutation is a deletion.

Accuracy: 0 %

Points: 0

Figure 1d. Identification Game (Balloon Bust)

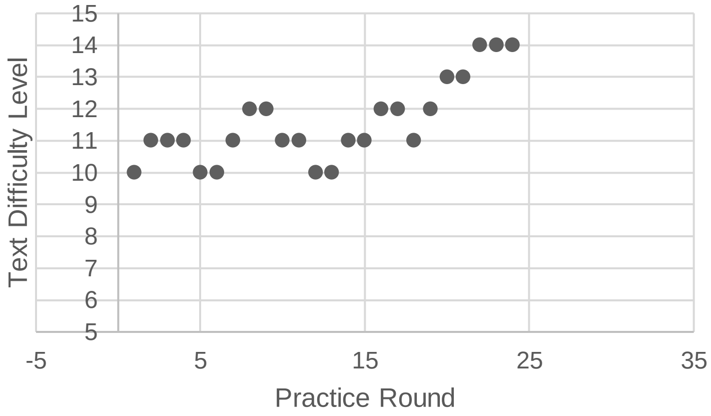


Figure 2. Text difficulty level over time for a participant with a high GMRT score (high reading skill) in the adaptive condition

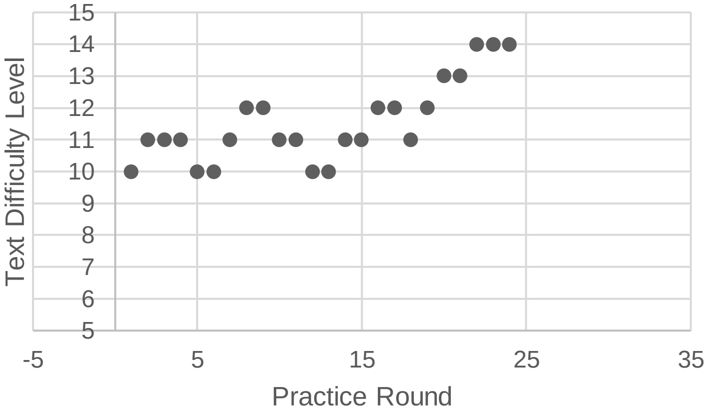


Figure 3. Text difficulty level over time for participant in random condition

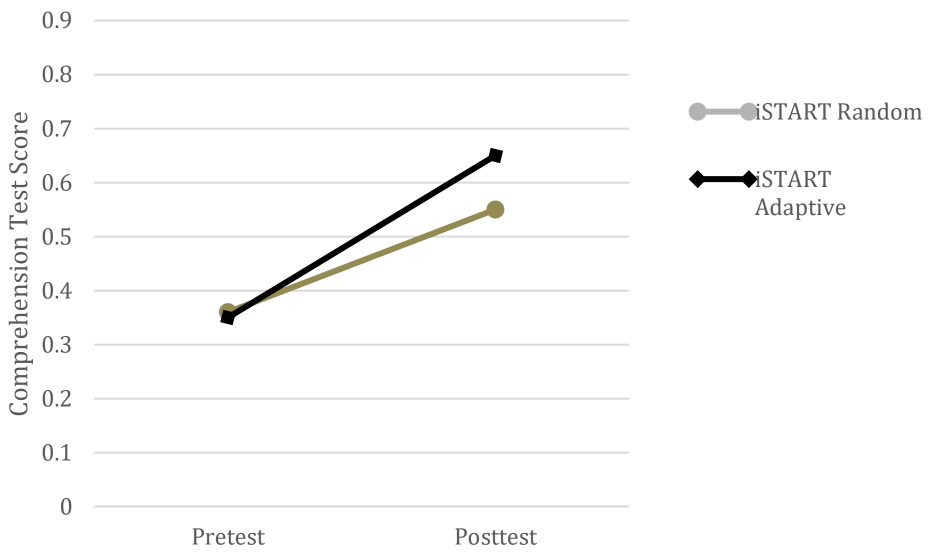


Figure 4. Comprehension Test Scores of Less Skilled Readers as a Function of Condition

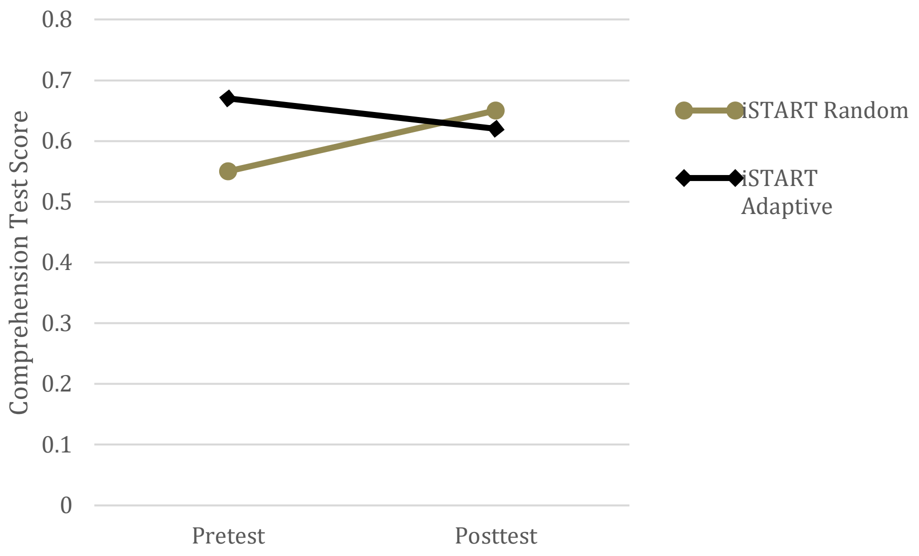


Figure 5. Comprehension Test Scores of Skilled Readers as a Function of Condition

Table 1.

Means and Standard Deviations of overall training experience [from (1) very bad to (6) very good] as a function of session and adaptivity condition

	Today's session was:	
	Random (n = 31)	Adaptive (n = 28)
	M (SD)	*M (SD)*
Session 1	4.10 (1.60)	4.43 (1.10)
Session 2	3.97 (1.30)	4.36 (1.13)
Session 3	4.13 (1.34)	4.43 (1.00)

Table 2.

Means and Standard Deviations of Negative Affect and Experiences [from (1) Never to (5) Always] as a function of session and adaptivity condition

	I was bored during today's session		**I was frustrated during today's session**		**I had problems with the program during today's session**	
	Random	Adaptive	Random	Adaptive	Random	Adaptive
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Session 1	3.19 (1.17)	3.07 (0.86)	2.61 (1.31)	2.61 (1.10)	2.00 (1.26)	2.18 (0.98)
Session 2	3.58 (.96)	3.14 (1.08)	2.32 (1.25)	2.50 (1.04)	2.06 (1.26)	2.07 (0.94)
Session 3	3.42 (1.03)	2.93 (1.09)	2.03(1.14)	2.32 (1.12)	1.68 (1.05)	2.32 (1.06)

Table 3.

Means and Standard Deviations of positive experiences [from (1) Not at all to (5) Very much] as a function of session and adaptivity condition

	I felt like I learned the material during today's session		**I feel like my reading skills improved during today's session**		**I enjoyed today's session**	
	Random	Adaptive	Random	Adaptive	Random	Adaptive
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Session 1	3.45 (1.06)	3.75 (0.65)	3.45 (1.31)	3.46 (1.14)	3.16 (1.44)	3.25 (1.30)
Session 2	2.94 (1.00)	3.64 (0.83)	2.77 (1.09)	3.32 (1.31)	2.77 (1.06)	3.21 (1.23)
Session 3	2.74 (0.97)	3.39 (0.88)	2.71 (1.24)	3.32 (1.02)	2.84 (1.16)	3.36 (0.95)

Table 4.

Gates-MacGinitie Reading Test, Self-Explanation, and Comprehension Scores as a Function of Training Condition

	GMRT	**Self-Explanation Scores**		**Comprehension Test Scores**	
		Pretest	Posttest	Pretest	Posttest
		M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
iSTART Random (n = 33)	.47 (.25)	1.64 (.48)	1.86 (.54)	.42 (.21)	.49 (.25)
iSTART Adaptive (n = 32)	.53 (.25)	1.66 (.46)	1.82 (.44)	.48 (.26)	.59 (.23)

Note. *t-tests compare pretest to posttest performance; *p < .05; + p < .10*

Table 5.

Correlations between reading skill, self-explanation scores, and comprehension test scores

(*n* = 65)	***M* (*SD*)**	**1**	**2**	**3**	**4**	**5**
1. GMRT (%)	0.51 (.25)	1.00				
2. Pretest Self-Explanation (0-3)	1.65 (47)	**0.24**	1.00			
3. Posttest Self-Explanation (0-3)	1.84 (.49)	**0.48**	**0.40**	1.00		
4. Pretest Comprehension Test (%)	0.45 (.25)	**0.52**	**0.49**	**0.43**	1.00	
5. Posttest Comprehension Test (%)	0.54 (.25)	**0.43**	**0.27**	**0.43**	**0.51**	1.00

Note. *Correlations significant at the p < .01 are indicated in bold*

Table 6.

Likelihood Ratio Tests for Predicting Self-Explanation Scores

Model	**Variables Added**	**AIC**	**BIC**	χ^2	***p***
M0	GMRT	2291.5	2311.8		
M1	+ Pretest/Posttest	2265.9	2291.2	27.60	> .001
M2	+ Training Condition	2267.5	2297.9	0.44	0.51
M3	+ Interactions	2254.9	2305.5	20.61	> .001

Table 7.

Summary of Linear Mixed Effects Model (M3) Analysis for Self-Explanation Scores

	B	***SE***	***t***	***p***
GMRT	0.32	0.07	4.73	0.00
Test (Pre, Post)	-0.24	0.05	-4.90	0.00
Condition (Random, Adaptive)	-0.10	0.10	-1.07	0.29
Test * Condition	0.09	0.07	1.30	0.19
Test * GMRT	-0.21	0.05	-4.29	0.00
Condition * GMRT	-0.17	0.10	-1.71	0.09
Test * Condition * GMRT	0.17	0.07	2.41	0.02

Table 8.

Likelihood Ratio Tests for Predicting Comprehension Test Scores

Model	**Variables Added**	**AIC**	**BIC**	χ^2	***p***
M0	GMRT	1202.70	-587.44		
M1	+ Pretest/Posttest	1198.70	-581.99	10.90	> 0.001
M2	+ Training Condition	1204.50	-581.42	1.13	0.29
M3	+ Interactions	1221.90	-576.23	10.38	0.03

Table 9.

Summary of Linear Mixed Effects Model Analysis (M3) for Comprehension Test Score

	B	***SE***	***t***	***p***
GMRT	0.13	0.04	3.62	0.00
Test (Pre, Post)	-0.07	0.04	-1.93	0.05
Condition (Random, Adaptive)	0.07	0.05	1.38	0.17
Test * Condition	-0.05	0.05	-0.98	0.33
Test * GMRT	-0.05	0.04	-1.30	0.19
Condition * GMRT	-0.07	0.05	-1.25	0.21
Test * Condition * GMRT	0.15	0.05	2.92	0.00