INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2022.2114143

iSTART: Adaptive Comprehension Strategy Training and Stealth Literacy Assessment

Danielle S. McNamara[a], Tracy Arner[a], Reese Butterfuss[a], Ying Fang[a], Micah Watanabe[a], Natalie Newton[a], Kathryn S. McCarthy[b], Laura K. Allen[c], and Rod D. Roscoe[d]

[a]Department of Psychology, Arizona State University, Tempe, AZ, USA; [b]Department of Learning Sciences, Georgia State University, Atlanta, GA, USA; [c]Department of Psychology, University of New Hampshire, Durham, NH, USA; [d]Human Systems Engineering, Arizona State University, Mesa, AZ, USA

ABSTRACT

The Interactive Strategy Training for Active Reading and Thinking (iSTART) game-based intelligent tutoring system (ITS) was developed with a foundation of comprehension theory and principles of learning science to improve students' comprehension of complex scientific texts. iSTART has been shown to improve reading comprehension for learners from middle school through adulthood, particularly lower knowledge readers, through strategy instruction and game-based practice. This paper describes iSTART, the theoretical foundations that have guided iSTART development, and evidence for the feasibility of game-based practice to improve learning outcomes. This paper also introduces a novel method of assessing students' reading comprehension through game-based literacy assessments that have been incorporated in iSTART. The development of these stealth assessments was guided by recent work emphasizing the need for rapid, dynamic, and low stakes assessments that evaluate students' reading skills in the context of brief, dynamic games. Stealth assessments can generate estimates of multiple aspects of students' reading comprehension quickly and within a motivating environment. The work described in this paper is a promising method to assess students' literacy in an unobtrusive and authentic way that may lead to improved learning outcomes for students.

1. Introduction

The Interactive Strategy Training for Active Reading and Thinking (iSTART) intelligent tutoring system (ITS) was developed to improve students' literacy skills. More specifically, iSTART leverages interactive and adaptive technology to enhance students' ability to comprehend the types of complex informational texts they encounter in their STEM courses. iSTART's development has been guided by theories of comprehension, learning, and human-computer interaction (HCI) to offer high-quality and effective comprehension strategy instruction and practice. Over the past two decades, advances in HCI and interactive computing have led to the continued refinement of the iSTART system.

This paper describes the theory of reading comprehension and principles in learning science that undergird the development of iSTART. Additionally, this paper includes how iSTART has been redesigned to incorporate technological advances that further support students' learning such as natural language processing (NLP), adaptivity, game-based practice, and novel, game-based (i.e., stealth) literacy assessments. These stealth assessments are incorporated into brief, efficient, engaging games to provide authentic, multidimensional information regarding students' literacy skills. The overarching goal in the development and redesign of iSTART is to address students' struggle with comprehension of complex, informational text.

1.1. The need to improve literacy

Although reading is an everyday activity, many students struggle to understand what they read, particularly when trying to learn from complex expository texts. Recent data from the Program for International Student Assessment (PISA) showed that U.S. adolescents ranked 15th in literacy skills worldwide. Similarly, roughly half of U.S. high school students do not have sufficient literacy skills to meet the demands of college-level reading (American College Testing, 2006). Insufficient literacy skills, then, become a barrier to students' access to and completion of post-secondary education (Baer et al., 2006). Of particular concern is the growing "STEM crisis," a shortage of individuals qualified to work in essential Science, Technology, Engineering, and Math (STEM) fields. STEM workers make up between 5% and 20% of workers in the U.S. (Xue & Larson, 2015) and that number is steadily growing. Under preparedness for college-level learning material is a key factor that contributes to low enrolment and success in STEM fields (Sithole et al., 2017),

CONTACT Danielle S. McNamara dsmcnamara1@gmail.com; Danielle.McNamara@asu.edu Department of Psychology, Arizona State University, 950 S. McAllister, Tempe, AZ 85287-1104, USA

increased STEM attrition rates, and dropping out of college altogether (Chen, 2013). Generally, the STEM-related jobs that are most in-demand require some level of college education (Fayer et al., 2017; Xue & Larson, 2015). Informational texts such as those that students encounter in STEM courses are particularly challenging.

Informational texts tend to be full of complex vocabulary and syntax and often overly rely on students' familiarity with or prior knowledge of the topic (Best et al., 2005; McNamara et al., 2011, 2012). Skilled readers enact comprehension strategies that can support their understanding. However, many students need more explicit strategy instruction and practice to build up their literacy skills. Thus, there is a great need to increase support for developing students' literacy skills to improve their success in both college-level coursework and subsequently, the future STEM workforce. iSTART leverages interactive technology to provide evidence-based comprehension strategy instruction and practice to develop students' literacy skills.

2. Comprehension theory underlying the development of iSTART

Successful comprehension of complex texts depends on the reader, the text, the task, and the context. This combination of factors requires the reader to coordinate several lower-level and higher-level processes to build a mental representation of what they read (McNamara & Magliano, 2009; C. Snow, 2002). One key process that distinguishes skilled and less skilled students is inferencing (Long et al., 1994; Oakhill, 1984). Inferences are necessary during reading to fill in information not explicitly stated in the text (Graesser et al., 1994; van den Broek et al., 2005). Therefore, students must generate inferences by connecting content they read with information they previously read or with their prior knowledge (Bowyer-Crane & Snowling, 2005; Elbro & Buch-Iversen, 2013; Kendeou, 2015; McNamara & Kintsch, 1996). However, students often fail to make inferences during reading (Cain & Oakhill, 1999; Cain et al., 2003) which leads to an impoverished understanding of the text's deeper meaning despite having knowledge of the text's explicit content. Thus, comprehension of a complex text requires that readers construct a coherent mental representation of the information, and they must generate inferences to do so (McNamara & Magliano, 2009).

The Construction-Integration (CI) Model (Kintsch, 1988, 1998) provides an account of the way in which readers construct meaning from texts at multiple levels of representation. According to the CI model, comprehension is the result of two cyclical phases. The first phase, construction, refers to the activation of information from the text and related prior knowledge. Knowledge activation stems from four sources: (1) the text information that is currently being processed, (2) the previously read sentence, (3) related prior knowledge, and (4) reinstatements from previously read text. The result of the construction phase is an associative network of new information gleaned from the text combined with the individual's prior knowledge that is activated during reading.

The second phase, *integration*, captures the spread of activation within this associative network. Highly connected information receives more activation than less connected information. Therefore, the more connections individuals make, the more likely information will be maintained in the network and will then become part of the reader's mental representation of the text. Conversely, information that has relatively few connections loses activation and is less likely to be maintained in the reader's final mental representation of the text (McNamara & McDaniel, 2004; Rowe & McNamara, 2008).

The CI model also assumes that the reader's final mental representation is multi-layered and includes a textbase (i.e., explicit in the text) representation and situation model (i.e., explicit information plus inferences beyond the text). Thus, a textbase-level understanding is relatively shallow and short-lived compared to a situation model. If students lack skill in drawing inferences to generate or retrieve information that was not provided in the text, then their situation model is likely to be relatively impoverished and lack coherence (McNamara & Kintsch, 1996; McNamara & McDaniel, 2004). When readers have less knowledge about a topic, they must use strategies to generate inferences and fill in conceptual gaps in the text using general world knowledge, reasoning, and common sense (McNamara, 2004).

2.1. The importance of comprehension strategy instruction

Comprehension strategies are explicit skills that support readers' construction and integration of new information as they develop situation models of text they are reading. Skilled readers automatically engage in active and elaborative comprehension strategies that help them to construct a coherent mental representation of the text (Coté et al., 1998; Goldman et al., 2012; Wolfe & Goldman, 2005). By contrast, less skilled and less knowledgeable students engage in these strategies less frequently, if at all. Providing strategy instruction encourages students to monitor their comprehension and engage in strategies that mimic those exhibited by skilled students. However, sustained training and deliberate practice of comprehension strategies not only strengthen comprehension skills among less skilled students, but they also encourage automatic use of those skills (McNamara, 2009).

Extensive evidence indicates the benefit of teaching students to self-explain text (Chi et al., 1994), generate deep-level questions (Davey & McBride, 1986), and summarize what they are reading (Brown et al., 1981). Self-explanation involves students explaining what they are reading to themselves using words that they are familiar with as well as connecting parts of the text and existing knowledge (McNamara, 2004). Within the framework of self-explanation instruction, students are first taught "comprehension monitoring" – how to recognize a breakdown in their understanding, which ideally encourages subsequent strategy use to improve understanding (Baker & Brown, 1984). One explicit strategy that aids in comprehension monitoring is question asking. Generating deep-level questions encourages students to

interrogate the meaning of text, their understanding of it as they read, and subsequently repair gaps in understanding (Davey & McBride, 1986; Rosenshine et al., 1996).

Self-explanation instruction supports students' comprehension by scaffolding lower level textbase models and guiding the development of higher-level situation models (see McNamara et al., 2007). At the sentence level, *paraphrasing* (i.e., restating texts in one's own words) helps students jumpstart deeper level comprehension processes. Paraphrasing aids in constructing a textbase-level representation of the text and in turn, improves memory of its main gist (McNamara et al., 2006). To go beyond the textbase and construct a coherent situation model of a text, students must also engage in knowledge-building strategies that elicit inferences (Scardamalia & Bereiter, 2006). Students can (1) "bridge" distal information in the text to form a more cohesive global understanding (Kintsch, 1998); (2) "elaborate" text content by drawing associations with prior knowledge, which in turn requires using logic and analogical reasoning to identify connections to previous experiences; and (3) "predict" subsequent text content based on their prior knowledge. Finally, summarization is a strategy that helps students to recognize the key information in text, by identifying what constitutes an extraneous detail and what constitutes an important main idea (Brown et al., 1981). Thus, summary generation supports the reader's development of a coherent mental representation of the text.

Training less skilled students to implement these strategies in the context of classroom instruction is effective (McNamara, 2004, 2017), albeit time consuming. Moreover, teachers must first understand how to model the strategies. These constraints limit the feasibility of successful strategy instruction as part of regular, direct classroom instruction. In the face of these challenges, automated comprehension strategy instruction can be delivered successfully due to the recent advances in interactive technologies.

2.2. Supporting comprehension with interactive technologies

Recent advances in interactive technology have made it possible to provide students with learning opportunities that mimic human instruction. Previously, important aspects of face-to-face instruction, such as feedback or personalization, were nearly impossible to provide without human intervention. Now, interactive technologies such as multimedia (e.g., video, audio, images), natural language processing (NLP), and artificial intelligence can support students' development of a coherent mental model.

Multimedia instruction supports the development of students' mental model during the construction phase using the "signaling" principle (Mayer, 2008, 2021) to notify the learner where there are opportunities to execute a comprehension strategy. For example, the ITS might signal students when they should generate a bridging inference to connect sentences in the text. Signaling can be executed using an image (e.g., arrow), a pop-up (e.g., dialog box with a text-entry field), or a prompt (e.g., audio notification) delivered via a pedagogical agent as in iSTART (Johnson et al., 2015; Wang et al., 2018). Interactive technologies can also support the integration phase of comprehension through explicit prompts to identify relevant prior knowledge related to the text the student is reading. Prompts may include generation of an elaborative inference – integrating information in the text with prior knowledge, or generation of a deep-level question about the connection between the text and world knowledge.

Beyond prompts to generate inferences, interactive technologies also facilitate evaluation of inferences that students generate during self-explanation (Bai et al., 2022). NLP combined with artificial intelligence can produce a score and provide automated feedback as well as monitoring students' performance and progress. Depending on the student's score, they may be guided to scaffolded strategy instruction that demonstrates the strategy being used followed by additional opportunities to practice and receive feedback.

Interactive technologies may also support students' development of a mental model by increasing students' engagement with the learning opportunity. The introduction of game-based learning and assessment provides learning opportunities that increase enjoyment of learning and motivation to engage in sustained practice (Jackson, Varner, et al., 2013).

3. Interactive strategy training for active reading and thinking (iSTART)

iSTART is an interactive technology developed to address the need to improve students' comprehension of complex scientific texts. iSTART provides adaptive, interactive strategy instruction and practice modeled after effective, human-based instruction implementing evidence-based principles from learning science (Jackson et al., 2015). The development of iSTART was guided by three well-known principles in the learning sciences: the generation effect, deliberate practice and feedback, and antidotes to disengagement to reduce learners' disengagement from the system (Healy et al., 2012; McNamara et al., 2015). Table 1 provides specific examples of how these foundational principles were instantiated in iSTART followed by descriptions of the key components in iSTART.

The first of these principles is the generation effect, following from the premise that individuals have improved recall when they actively generate content compared to when they read or copy content verbatim (McNamara & Healy, 1995). For example, when students generate a target word from a word pair cue, they will have improved subsequent recall of that word pair compared to when they simply read it. iSTART executes this principle by eliciting constructed (i.e., generated) responses in practice activities and mini-games.

The second principle guiding iSTART is deliberate practice with feedback. For practice to be effective, it must be deliberate and effortful, and learners must be motivated to improve on a targeted weakness (Ericsson, 2008). Providing feedback during deliberate practice informs the learner of

Table 1. Foundational principles from learning science used in the development of iSTART.

Foundational principle	Explanation	iSTART Example
Generation effect	Individuals have better memory for content that they generate than content that they read or write verbatim.	Learners generate self-explanations of target sentences during practice and mini games.
Deliberate practice and feedback	Deliberate practice is effortful, sustained, focused on a specific learning target, includes feedback, and is executed by a motivated learner.	Quality scores and feedback are provided to learners during practice through the use of NLP algorithms.
Antidotes to disengagement	Deliberate practice must be sustained to be effective, thus learners must remain engaged in the practice activity.	Mini-games in iSTART increase learner agency and personalization which deter disengagement with the system.

progress towards their goal, what needs to be improved, and how to improve it. iSTART instantiates deliberate practice and feedback in both the training and practice portions of the system. For example, students practice generating self-explanations after which they receive a quality score driven by an NLP algorithm that uses latent semantic analysis and word-based measures. Students receive feedback based on their self-explanation quality score which may include positive affirmations (e.g., "good job") for higher scores and prompts to revise for lower scores (e.g., "try writing a longer self-explanation").

The generation effect, deliberate practice, and feedback have an abundance of research demonstrating their benefits to students' learning. However, students can easily lose motivation without additional incentives to persist, especially when confronted with challenging tasks (Jackson & McNamara, 2017). To that end, iSTART includes features that serve as "antidotes to disengagement" (e.g., mini-games and personalization) to motivate students to remain engaged in deliberate practice and increase agency (Jackson & McNamara, 2013).

3.1. Key components of iSTART

3.1.1. Comprehension strategy video lessons

iSTART instruction is divided into three modules introducing active reading comprehension strategies: Question Asking, Summarization, and Self-explanation. The format of instruction in each module includes an overview of the strategy, instruction on how to execute the strategy, and a recap of strategy use. The instruction videos model human instruction and follow the explicit direct instruction ("I do, we do, you do") method (Rupley et al., 2009) encouraging students to work along with a narrator. First, the narrator defines and demonstrates how to use the strategy. Then, the narrator verbally models using the strategy and suggests that the student does the same while watching the modeling. Finally, students are provided with an example and instructed to pause the video and try the strategy on their own using a sentence on the screen. Following the "instruction pause," the narrator provides possible solutions to the student and directs them to "check their answer." At the end of the video, students complete a short quiz to assess their understanding of the strategy.

The Question Asking module supports comprehension monitoring by prompting students to identify parts of text that they do not understand and then searching for the answer in the text (Afflerbach et al., 2020; Oakhill et al., 2019). The lessons in this model also demonstrate the types of questions students can ask themselves to ensure that they understand the larger meaning of the text. The Summarization module provides information about how to generate high-quality text summaries, such as deleting unimportant details, identifying main ideas, replacing portions of text with words that are easier to understand, and identifying the topic sentences that indicate the overall meaning of the text (Stevens et al., 2019). Students are taught strategies to generate inferences in the Self-explanation module. Self-explanation training is divided into five strategies that support comprehension: comprehension monitoring, paraphrasing, bridging and elaborative inferences, and prediction. Generating self-explanations increases students' use of bridging and elaborative inferences, which lead to the construction of a more coherent situation model of the text (Kintsch, 1998; McNamara, 2004, 2017).

3.1.2. Strategy practice

Following the video lessons, students are provided opportunities to practice the strategies through coached practice or game-based practice. During both forms of practice, students are provided feedback that is delivered through natural language processing (NLP) based algorithms (Crossley et al., 2016; Johnson et al., 2017; McCarthy, Allen, et al., 2020). These algorithms generate both formative and summative feedback that simulate what students might receive from teachers or peers in the classroom. Summative feedback is provided in the form of a score (i.e., poor, fair, good, or great) and formative feedback is given in the form of personalized suggestions for improving the quality of self-explanations. iSTART further scaffolds students' strategy learning using NLP based algorithms to adaptively select texts with an appropriate level of difficulty based on each student's skills. Formative and summative feedback is integrated throughout iSTART to guide students to components and activities that will address knowledge gaps as they proceed through the system.

Coached practice includes scaffolded support provided by Mr. Evans, a pedagogical agent, who gives students explicit feedback about their practice. Students receive feedback indicating that they have done well (e.g., *Great job!*) or that they need to try again (e.g., *Try making your explanation longer.*). Students may also receive guidance on how to improve on the next sentence (e.g., *Next time, try to improve the explanation by including information from different parts of the text.*)

Game-based practice is divided into two primary types: identification games and generative games. In identification games, students view example responses (e.g., questions, summaries, self-explanations) and must correctly identify the strategy used. Whereas, in generative iSTART games, students practice generating questions, summaries, or self-explanations of target sentences.

3.1.3. Antidotes to disengagement

iSTART includes systemwide game elements common to both recreational and educational game platforms to increase agency and motivate students to continue playing (Seaborn & Fels, 2015). These features include game-based practice and personalization features in the interface and instruction.

Game-based Practice. Students' motivation to engage in practice is particularly bolstered by the incorporation of strategy practice mini-games (Jackson & McNamara, 2013, 2017; Jacovina et al., 2016). A wide variety of mini-games offers students high levels of agency in choosing which game to play, or even whether to play a game or engage in coached practice (E. L. Snow et al., 2014). Players earn points during game play, which accumulate to earn bronze, silver, or gold trophies. Students also earn trophies for reaching new high scores, which can then be used to open new game levels or "level-up." The thrill of attaining the next level can provide the motivation that students need to continue playing when faced with difficult content (Tsai et al., 2020). These systemwide gamification features support learning by enticing students to engage in extended practice (Jackson & McNamara, 2013, 2017).

Mini-games in iSTART were originally developed to support students' acquisition of comprehension strategy knowledge (i.e., strategy identification games) and to provide engaging, sustained practice using the strategy (i.e., generation games; McCarthy et al., 2020).

One example of an identification game is Balloon Bust (Figure 1). Students are presented with a text and an example self-explanation and their task is to "throw" a dart at moving balloons that depict the strategy used to generate the self-explanation. Students are challenged further when a self-explanation demonstrates two strategies, and they must identify them both and bust the balloons to move on. Students receive points for accurate responses, lose points for inaccurate responses, and get bonus points for not making any mistakes (i.e., selecting the wrong strategy). At the end of a round, students can view their score summary before going on to the next round. Identifying the strategies used in games like Balloon Bust can aid students' understanding of how the strategy is applied through multiple observations of each one (Jackson & McNamara, 2013). Self-Explanation Showdown is an example of a generation mini-game in which students compete against a computer opponent in a gameshow style scenario. The student and an opponent each generate a self-explanation for the same sentence while NLP algorithms evaluate them and assign a summative score (Jackson et al., 2012). The player with the higher score gets points for winning the round. Students get to practice generating self-explanations

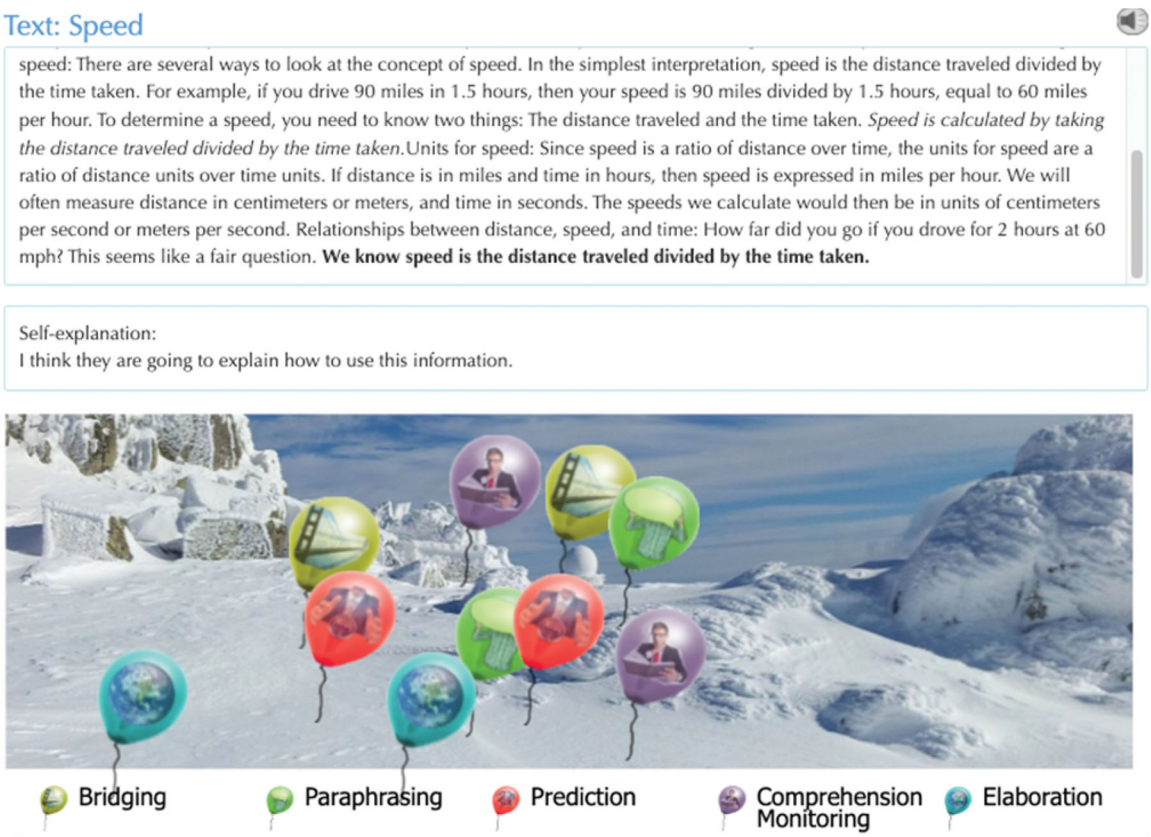

Figure 1. Balloon Bust strategy identification game.

while also getting the thrill of "beating the competition" and earning points.

More recently, additional types of game-based practice, beyond identification and generation, have been added to iSTART to better represent the diversity of comprehension assessment tasks that students complete throughout their educational careers. For instance, students' comprehension is often assessed using multiple-choice tests such as standardized assessments (e.g., ACT, SAT, NAEP). StairStepper is an adaptive game that combines comprehension assessment and self-explanation generation to support students' text comprehension and development of a situation model of the text. Students are able to monitor their comprehension of a text by responding to multiple-choice questions and receiving a score. Students who score below the threshold receive a comprehension scaffold in the form of a prompt to self-explanation target sentences in the text (Figure 2; Arner et al., 2021; Perret et al., 2017). The goal of StairStepper is for the player to move their avatar to the top of the staircase. Correct answers result in the avatar moving up and the student advancing to a more difficult text. Incorrect responses trigger a comprehension strategy, self-explanation, to help the students update or improve their mental model. Multiple incorrect responses cause the avatar to descend the stairs and students receive an easier text. Throughout the game, students receive adaptive feedback from Mr. Evans, a pedagogical agent, depending on the game level and their progress. When students are prompted to write self-explanations, they may receive feedback on writing a good self-explanation. When they achieve a new level and move up a step, they receive positive reinforcement (e.g., "Great job!"). The adaptive text and feedback further support students' persistence, leading to increased strategy use and improved reading comprehension (Arner et al., 2021; Jackson & McNamara, 2013).

Lost in Springdale (Figure 3) is a narrative game developed in a "Choose your own Adventure" format aimed at improving functional comprehension skills of adult readers such that they require players to solve real-world problems presented in text (e.g., reading clues to determine a distance on a map; Johnson et al., 2016). Readers navigate through Springdale, a town where many of the inhabitants have disappeared, assumedly because it was affected by a mysterious disaster. Readers are presented with the end goal of surviving the disaster, while also searching for clues to uncover the whereabouts of the townspeople. As readers proceed through stops on a map, they encounter real-world artifacts and must apply their reading comprehension skills (i.e., paraphrasing a clue) to proceed. Clues are modeled after three types of text literacy: prose literacy (e.g., news articles), document literacy (e.g., nutrition labels), and quantitative literacy (e.g., calculating a distance on a map) identified by the National Assessment of Adult Literacy. Lower literacy students who may have insufficient decoding skills are supported by pedagogical agents who read the text aloud while it is presented on the screen (Johnson et al., 2017). Readers solve each clue by asking or answering questions and generating self-explanations and summaries. Throughout the game, readers may also access an in-game cell phone that includes features like photographing key areas on the map, note taking, and skills trackers that show readers their progress on comprehension skills (e.g., summarization) and concept knowledge (e.g., health). Points are awarded for each successful task based on the complexity of the task, and readers' progress through the narrative, which can then be used to "upgrade" their cell phone. Lost in Springdale combines adaptive story elements, life-relevant task artifacts, and game elements to support sustained practice of reading comprehension strategies. While this game was designed for low literacy adults, the engaging and flexible design has strong potential to also benefit school-age learners (Johnson et al., 2017).

Personalization in iSTART. iSTART includes features that afford personalization by both the student and the teacher. Students can personalize the iSTART interface using in-game currency, iBucks, earned during practice. Thus, engaging in more practice increases the student's ability to personalize the system. Personalization options include changing the

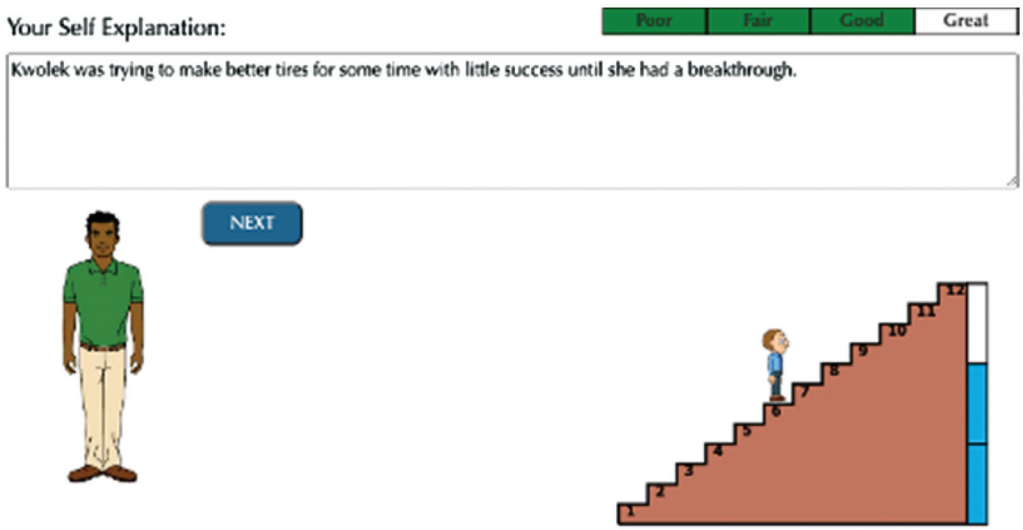

Figure 2. StairStepper self-explanation game.

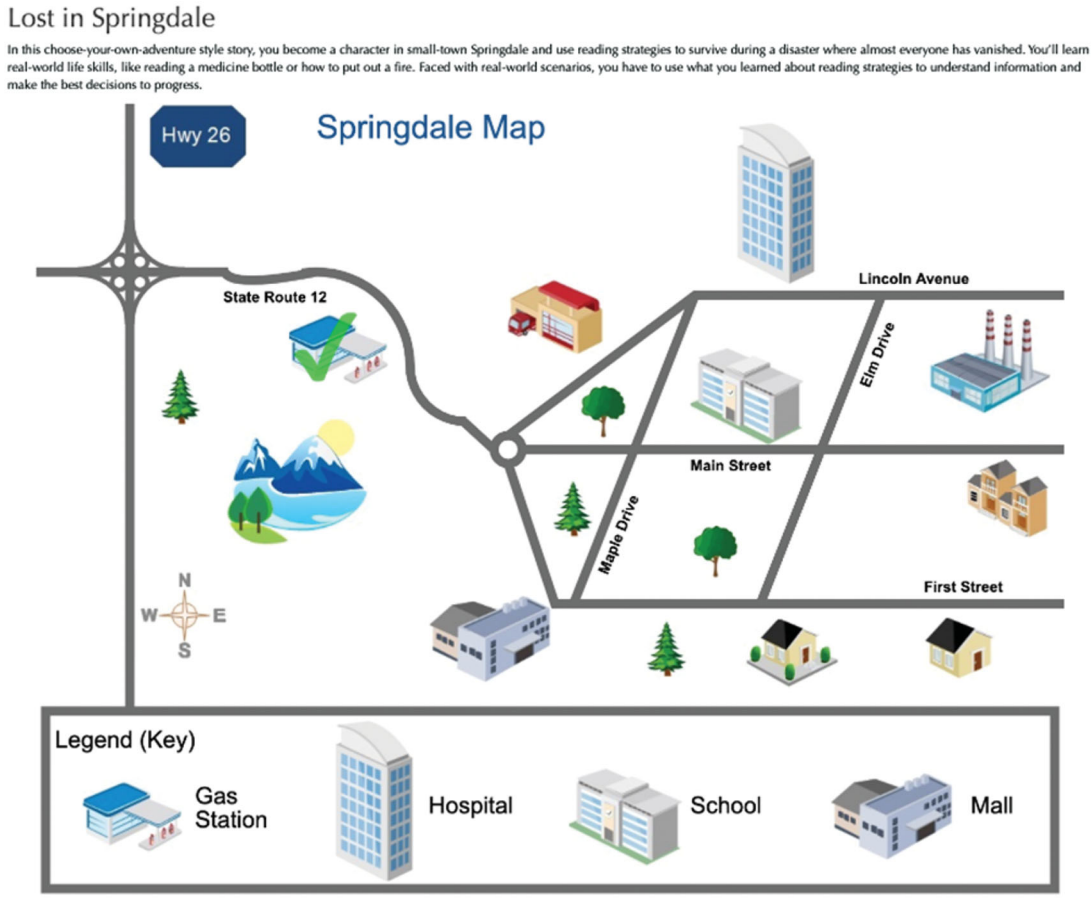

Figure 3. Lost in Springdale interactive narrative game.

appearance of the player's avatar or editing the background color of the interface. Students who interacted more with the games and the game-based features reported higher levels of engagement and motivation and showed higher performance in terms of generating better quality self-explanations (E. L. Snow et al., 2013a, 2013b).

iSTART can also be personalized by the teachers via an interface that allows customization of their iSTART classroom. The teacher dashboard includes settings for lesson order, initial text difficulty, whether students can play games, and when they can do so. For example, teachers can set prerequisites to open each strategy training such as the passing level for checkpoint quizzes. Teachers can also select specific games for students to play and set the starting text level for each game. The text library in iSTART allows teachers to assign specific texts that align with their curriculum (e.g., Newton's laws of gravity), which may support students' concept learning through practicing comprehension strategies.

3.2. Evaluating the effectiveness of iSTART

It is critical to evaluate the efficacy of any intervention with diverse groups of learners, thus establishing which interventions work, for whom, and under what conditions. Despite evidence that reading skill acquisition requires prolonged, deliberate practice, many interventions are often tested in brief laboratory studies with convenience samples (e.g., undergraduate research pool). Intervention effectiveness may not be accurately evaluated in studies with shorter dosages, particularly with learners at varying levels of knowledge or reading skill. To that end, the effectiveness of iSTART has been experimentally tested with middle-school and high-school students (McCarthy et al., 2020; McNamara et al., 2006), college students (Arner et al., 2021), and adult literacy students (Johnson et al., 2017) in both shorter and longer intervention durations.

The combination of active reading comprehension strategy instruction, game-based practice, and feedback in iSTART has been shown to increase students' engagement and motivation to practice and improve reading comprehension in several learner populations (Jackson et al., 2015; McNamara, 2017; McNamara et al., 2006). One such study evaluated the effectiveness of iSTART with middle school students. The experimental group completed approximately two hours of iSTART strategy training compared to a control group who received an overview of the strategies in iSTART but no training. Students in the training condition produced more elaborations in their self-explanations than did students in the control condition. The training was beneficial for students with both higher and lower strategy knowledge but, students who had greater prior knowledge of reading strategies produced better quality self-explanations after this short intervention (McNamara et al., 2006). These results suggest that longer training durations may improve learning outcomes for students with lower reading skill.

To this end, Jackson, Varner, et al. (2013) investigated the benefit of increasing the dosage of iSTART training to eight 1-hour sessions with high school students. Indeed, results indicated that less-skilled students showed a greater increase in the quality of their self-explanations compared to more-skilled students. While, more-skilled students did benefit from the intervention, their gains were smaller than those of less-skilled students. One important finding from this work is that students who reported the lowest commitment to reading (i.e., minimal independent reading practice) demonstrated improved self-explanation quality from pretest to posttest (Jackson, Varner, et al., 2013). Similarly, a study conducted over the course of a school year (6 months) demonstrated that both skilled and less-skilled students improved their self-explanations over time (Jackson et al., 2010). However, in the longer intervention, less-skilled students improved more than the more-skilled students, and the abilities of skilled and less-skilled students converged by the end of the school year. One possibility for this convergence is that the more-skilled readers were automatically engaging in active reading strategies whereas the less-skilled readers did not know how to use strategies prior to the study (Jackson et al., 2010). In sum, the results of these studies show that learners across grade levels benefit from iSTART training. However, the specific skills gained by learners vary by ability level indicating that further research is needed to clarify what features of iSTART are most beneficial for both more-skilled and less-skilled readers.

iSTART has undergone several modifications in the past decade including the addition of training on Question Asking and Summarization strategies, adaptive text selection, improving personalized feedback, and the use of stealth assessment through gameplay (McCarthy et al., 2020). Further research is needed to assess the impact of these added components, as well as how to adapt instruction to meet the needs of each individual student and groups of students in classrooms. Key to adaptive instruction is the use of stealth assessment during instruction.

4. Stealth assessment

Stealth formative assessments are embedded in digital games to measure student knowledge and skills covertly and unobtrusively (Shute & Ventura, 2013). In stealth assessment, evaluation occurs during the activity rather than at the conclusion or by interrupting learners with separate tests or quizzes. Testing items are replaced with gaming tasks and activities such that learners are largely unaware of being evaluated. When students perform game tasks, they naturally produce rich sequences of actions and performance that become the evidence needed for knowledge and skills assessment.

In a well-designed game assessment scenario, students may not be aware that skills are being assessed. The experience of "play" has the potential to induce a greater sense of salience than the experiences common to "testing," which can render the assessments more natural and authentic. Such assessments are more engaging, more satisfying, are taken more seriously by learners, and can contribute to formative assessment and skill development (Gulikers et al., 2008; James & Casidy, 2018; Sotiriadou et al., 2020). Importantly, stealth assessments are based on students' learning behaviors (e.g., game-play actions and performance) rather than post-hoc measurements of performance. Moment-to-moment learning data can be used to assess students dynamically while capturing complex cognitive processes. Additionally, dynamic assessments can inform software learning environments about changes in students' abilities and skills, and these systems can subsequently adapt to students based on their specific pedagogical needs (VanLehn, 2006).

Stealth assessments have been integrated into a variety of game settings to evaluate players' skills and knowledge. For instance, in Use Your Brainz – a slightly modified version of a popular commercial game Plants vs. Zombies 2 – stealth assessments are embedded to evaluate problem-solving skills (Shute et al., 2016). While playing, students produce a dense stream of performance data that are recorded in log files and analyzed to infer students' problem-solving skills. Similarly, Physics Playground is a computer game based on 2 D physics simulations in which players guide a green ball from a predetermined starting point to a red balloon by drawing simple machines such as ramps, levers, pendulums, and springboards (Shute & Ventura, 2013). The performance-based stealth assessments used in these games have been validated against external measures of related skill and they meaningfully capture students' physics knowledge, persistence, and creativity (Shute & Rahimi, 2020; Shute et al., 2013; Ventura & Shute, 2013). Students' gameplay and performance data recorded in iSTART can function as stealth assessments of students' reading skill.

4.1. Stealth assessment in iSTART

In the current instantiation of iSTART, students' in-game performance is primarily used to provide feedback, (immediate and adaptive) and to assign texts appropriate to the students' estimated reading ability. However, iSTART's game-based practice serves as an ideal environment for stealth assessment of comprehension skills (Jackson, Snow, et al., 2013). In identification games, students' answer selections serve to diagnose students' understanding or confusion. In generative games, students' self-explanations are analyzed for evidence of strategy use. The following section describes current work expanding the use of stealth assessment in iSTART to dynamic assessment of students' literacy skills.

4.2. Assessment leveraging identification games

iSTART game-based practice includes comprehension strategy identification games that require multiple-choice responses. Thus, one approach to gleaning information about readers' literacy skills is to assess their performance on identification tasks that require simple point-and-click decisions during reading. Butterfuss et al. (2021) found that readers' accuracy

in selecting the most important sentences in expository text was positively related to their literacy skills. In this study, each sentence from the text was assigned an importance score that reflected the proportion of readers in the sample who chose the sentence as a main idea. Thus, sentences that were more frequently selected as important were assigned higher importance scores than sentences that were selected less frequently. Readers' mean importance scores for the sentences were significantly correlated with their comprehension skill ($r = 0.43$) and vocabulary knowledge ($r = 0.70$), as measured by the Gates-MacGinitie Reading and Vocabulary Tests (MacGinitie & MacGinitie, 1989). These results suggest that tasking readers with choosing sentences that include the most important information in the text during their engagement with iSTART may provide a proxy of literacy skills. Compared to traditional assessments of literacy skills, main idea selection tasks, are relatively easy to implement, can be embedded into virtually any academic text, and have potential to inform both formative feedback that guides readers' subsequent activity and summative feedback (i.e., a score) in the context of iSTART.

Fang et al. (2021) examined performance on three identification games to assess reading skills, as well as how players' attitudes towards the games was related to game performance and reading skills. Adult readers played three different identification games in iSTART. The first, Vocab Flash, is an adaptive vocabulary game in which readers are given a target word and asked to select a synonym from four alternatives (Figure 4). The difficulty of the vocabulary words adapts to readers' performance such that high-performing readers receive increasingly more difficult vocabulary words and low-performing readers receive easier words. This adaptivity mirrors computer-adaptive testing (Meijer & Nering, 1999) in which individuals can fluctuate between levels of difficulty according to their performance, with more skilled individuals encountering more difficult items and less skilled individuals encountering easier items. The second game was Dungeon Escape, which tasks readers to select the best topic sentence of a given text. The third game, Adventure's Loot, tasks readers to identify main ideas in a given text and avoid choosing unimportant information. Readers completed a brief measure of their perceptions of the game following each one (e.g., "This game was fun to play."). Finally, readers completed an online version of the Gates-MacGinitie Reading Test to measure their vocabulary and reading comprehension skill (4th ed., MacGinitie & MacGinitie, 1989).

As shown in Table 2, both game performance and perceptions of the games were significant predictors of participants' vocabulary and reading comprehension scores. Readers' performance on Vocab Flash accounted for a substantial amount of variance in both vocabulary knowledge and reading skill performance (~75%). The extent to which participants enjoyed the game did not account for additional variance. Thus, Vocab Flash successfully captured a substantial portion of the variance in literacy skills with only 5 minutes of game play. The two games that involved reading and choosing sentences, Dungeon Escape and Adventure's Loot, accounted for significant amounts of variance in both vocabulary knowledge and reading skill (19–25%), albeit substantially less than did Vocab Flash. Additionally, participants tended to enjoy the games; however, less skilled readers enjoyed the games more than skilled readers, and the degree to which they enjoyed the game accounted for additional variance in literacy skills. Apparently, less skilled readers found the games more challenging, which may have increased their engagement and in turn, the predictive power of their game performance. These results point to the importance of matching students to the appropriate level of challenge within learning games, which is more difficult in games where students read text as opposed to games in which students choose vocabulary words.

4.3. Assessment leveraging generation games

Another approach toward stealth literacy assessment in the context of iSTART leverages students' constructed responses

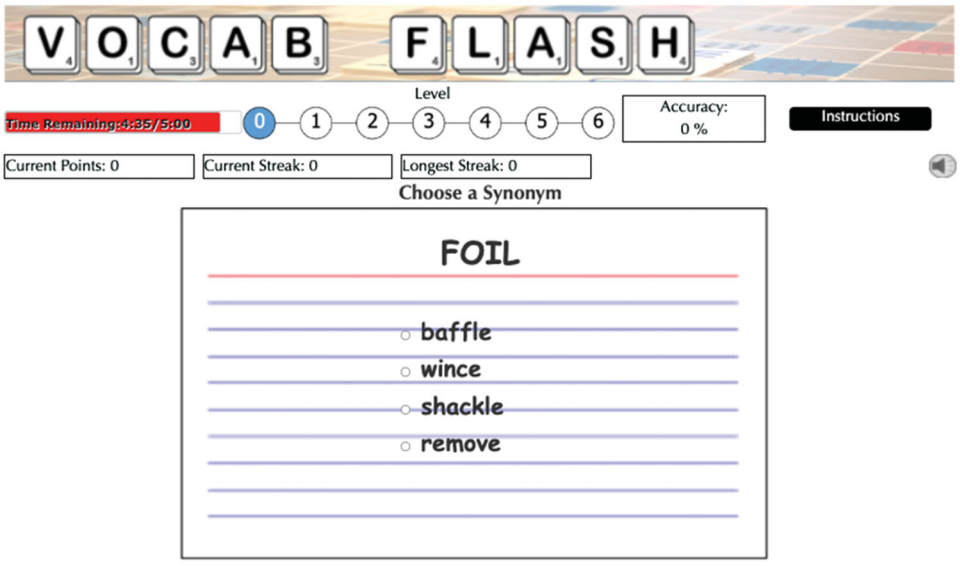

Figure 4. Vocab Flash adaptive vocabulary game.

Table 2. Proportion of variance accounted for by iSTART game performance and game enjoyment (from Fang et al., 2021).

	Vocabulary		Reading skill	
Game	Performance	Enjoyment	Performance	Enjoyment
Vocab flash	.76	.00	.74	.00
Dungeon escape	.20	.06	.25	.02
Adventurer's loot	.21	.24	.19	.17

(e.g., explanations and summaries) generated during practice or generative games. Prior work has shown that linguistic properties of constructed responses, such as essays, can be used to make inferences about students' literacy skills (e.g., Allen & McNamara, 2015). More specific to the types of constructed responses in iSTART (e.g., self-explanations), Allen et al. (2015) asked students to generate self-explanations while reading. They used NLP to calculate a set of descriptive (e.g., word count and average word length), lexical, syntactic, and cohesive indices of students' self-explanations. Twenty-four indices were significantly related to students' scores on a standardized reading test (i.e., Gates-MacGinitie Reading Test). Three linguistic features (lexical diversity, semantic cohesion, and sentence length) accounted for 38% of the variance in students' comprehension of science texts. This finding suggests that better readers tended to use a greater diversity of words and shorter sentences while connecting sentences in their self-explanations together through semantic ties.

Fang et al. (2021) investigated the feasibility of using responses generated during iSTART gameplay as stealth assessment of reading comprehension skill by analyzing data from two studies (McCarthy et al., 2018; McCarthy et al., 2020). Fang and colleagues (2021) aggregated self-explanations to create 12 *aggregated self-explanations* for each student, each of which included between 1 and 12 self-explanations. Across both datasets, the results indicated that the power of the linguistic features of self-explanations to predict reading comprehension skill increased (i.e., from 10% to 39% of the variance) as more self-explanations were included in the model. However, the increase was not statistically significant after including 9 self-explanations.

Collectively, these findings indicate that linguistic features of students' self-explanations successfully align with reading comprehension skills as measured by a standardized reading test. Thus, tasking students to produce a relatively modest number of self-explanations (i.e., ∼9) may provide an efficient means of evaluating their literacy skills. Consequently, game-based learning in iSTART provides an antidote to disengagement such that students are inclined to sustain deliberate practice and receive feedback for durations that both support acquisition of comprehension strategies and provide a mechanism to covertly assess learning (Fang et al., 2021; Jackson & McNamara, 2013, 2017).

5. Conclusion

The efficient and effective development of students' reading comprehension skills is an essential component of their academic, professional, and personal success; yet supporting such growth has remained a significant challenge for educators. iSTART was developed to address this challenge with a foundation built on the Construction-Integration model of reading comprehension (Kintsch, 1988) and three evidence-based principles in learning science: the generation effect, deliberate practice and feedback, and antidotes to disengagement. Over several decades of research and testing, the iSTART tutoring system has emerged as an evidence-based tool to improve reading comprehension strategies and learning outcomes for both adolescent and adult learners (Jacovina et al., 2016; Johnson et al., 2017; McNamara, 2006, 2017). Ongoing enhancements of iSTART have incorporated a variety of adaptive and motivating features that have further strengthened the utility and usefulness of the tutor, such as matching text difficulty to learners' reading skills, a gamified interface for personalization, and multiple options and forms of game-based practice (Jackson & McNamara, 2013; Johnson et al., 2016; McCarthy et al., 2020). The latter innovations have, in turn, empowered further explorations of game-based and stealth assessments (Butterfuss et al., 2021; Fang et al., 2021).

Game-based assessments enable measurement of students' performance and growth in a more engaging format, and stealth assessment can minimize both the salience and fatigue of testing – in some cases, students may not even know that they are being assessed. The ability to downplay the experience of learning and assessment through game-based interactions has strong potential to enable more authentic and dynamic estimations of students' abilities. This is a notable finding because these games are fast and simple (i.e., players do not need to generate text and NLP algorithms are not required), and yet still revealing (Fang et al., 2021). They imply exciting flexibility for future assessments that can be deployed in a number of game-based scenarios. In more complex games (i.e., with more player choices or agency), more complex models of reading skill may be possible – estimation accuracy should only increase with additional data. Additionally, work presented here demonstrated how generative games and coached practice, which require learners to compose self-explanations, can be mined to estimate reading skills. NLP tools and techniques can extract lexical, syntactic, structural, and semantic features of students' constructed responses to predict reading scores and performance. In turn, these assessments can drive feedback to help learners iteratively improve.

It is important to note that this work does not suggest that all learning activities, assessments, or standardized testing be replaced with games or stealth assessment. The benefit of instructional modalities (e.g., multimedia, delivery by a pedagogical agent) vary across learner populations and skill sets and different modalities for evaluation have different strengths and weaknesses. For example, the emphasis on "play" in game-based settings may lead learners to make decisions or take actions that align more with "fun" than with attempts to "win" or "perform." Students' playful attempts to test the limits or mechanics of a game (e.g., attempting to crash a spaceship or pop as many balloons as possible) might result in reduced learning or faulty estimates of their skills

because students are not *trying* to succeed as game designers imagined. Game-based instruction and stealth assessments must take into account this potential divergence. Nonetheless, the potential benefits (e.g., higher engagement, unobtrusiveness, and lower test anxiety) remain worthwhile.

Games offer insights and options that are not possible with in-person instruction, traditional assessments of learning, or standardized testing, particularly with regard to capturing learners' online behaviors as they practice reading comprehension strategies and work on acquiring new knowledge (Jackson, Snow, et al., 2013). Insights into students' comprehension through stealth assessments allows for real-time adaptation to meet the needs of the individual learner. For example, stealth assessment of comprehension skill can help scaffold learning for students with less concept knowledge by avoiding complex or knowledge-dependent question wording. Game-based practice and assessment can reduce time pressure or testing salience that may cause learners to experience high anxiety (von der Embse et al., 2018) while encouraging sustained, deliberate practice for learners with a variety of skill levels.

The research described in this paper is a promising indication that the engaging, game-based practice activities integrated into iSTART are also successful indicators of students' skill acquisition through stealth assessment. However, additional research is needed to determine what features of the system are most beneficial for different types of learners, and in particular, the types of online behaviors and system interactions that can support the development of learners' reading comprehension of complex or scientific text.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research reported here was supported by the Office of Naval Research, through Grants [N00014-17-1-2300, N00014-20-1-2623, N00014-19-1-2424, and N00014-20-1-2627] and the Institute of Education Sciences, U.S. Department of Education, through Grants [R305A190050 and R305A190063] to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Office of Naval Research, the Institute, or the U.S. Department of Education.

References

Afflerbach, P., Hurt, M., & Cho, B. Y. (2020). Reading comprehension strategy instruction. In D. L. Dinsmore, L. K. Fryer, & M. M. Parkinson (Eds.), *Handbook of strategies and strategic processing* (pp.98–118). Routledge.

Allen, L. K., & McNamara, D. S. (2015). You are your words: Modeling students' vocabulary knowledge with natural language processing. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)* (pp.258–265). International Educational Data Mining Society.

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. Baron, G. Lynch, N. Maziarz, P. Blikstein, A. Merceron, & G. Siemens (Eds.), *Proceedings of the 5th International Learning Analytics & Knowledge Conference (LAK'15)* (pp. 246–254). ACM.

American College Testing. (2006). ACT high school profile report: The graduating class of 2006. https://www.act.org/content/dam/act/unsecured/documents/Natl-Scores-2006-National2006.pdf

Arner, T., McCarthy, K. S., & McNamara, D. S. (2021). iSTART StairStepper—Using comprehension strategy training to game the test. *Computers*, *10*(4), 48. https://doi.org/10.3390/computers10040048

Baer, J. D., Cook, A. L., & Baldi, S. (2006). *The literacy of America's college students*. American Institutes for Research.

Bai, C., Yang, J., & Tang, Y. (2022). Embedding self-explanation prompts to support learning via instructional video. *Instructional Science*, 1–21. https://doi.org/10.1007/s11251-022-09587-4

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 353–394). Longman.

Best, R. M., Rowe, M., Ozura, Y., & McNamara, D. S. (2005). Deep-level comprehension of science texts: The role of the reader and the text. *Topics in Language Disorders*, *25*, 65–83.

Bowyer-Crane, C., & Snowling, M. J. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *The British Journal of Educational Psychology*, *75*(Pt 2), 189–201.

Brown, A. L., Campione, J. C., & Day, J. D. (1981). Learning to learn: On training students to learn from texts. *Educational Researcher*, *10*(2), 14–21. https://doi.org/10.3102/0013189X010002014

Butterfuss, R., Orcutt, E., Fang, Y., Kendeou, P., & McNamara, D. S. (2021, April 22–25). You pick'em: Selecting main ideas versus deleting details [Conference presentation]. American Educational Research Association (AERA) 2021 Annual Meeting.

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing*, *11*(5/6), 489–503. https://doi.org/10.1023/A:1008084120205

Cain, K., Oakhill, J. V., & Elbro, C. (2003). The ability to learn new word meanings from context by school-age children with and without language comprehension difficulties. *Journal of Child Language*, *30*(3), 681–694. https://doi.org/10.1017/S0305000903005713

Chen, X. (2013). *STEM attrition: College students' paths into and out of STEM fields. Statistical analysis report* [NCES 2014-001]. National Center for Education Statistics.

Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. https://doi.org/10.1207/s15516709cog1803_3

Coté, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes*, *25*(1), 1–53. https://doi.org/10.1080/01638539809545019

Crossley, S., Kyle, K., Davenport, J., & McNamara, D. S. (2016). *Automatic assessment of constructed response data in a chemistry tutor*. International Educational Data Mining Society.

Davey, B., & McBride, S. (1986). Effects of question-generation training on reading comprehension. *Journal of Educational Psychology*, *78*(4), 256–262. https://doi.org/10.1037/0022-0663.78.4.256

Elbro, C., & Buch-Iversen, I. (2013). Activation of background knowledge for inference making: Effects on reading comprehension. *Scientific Studies of Reading*, *17*(6), 435–452. https://doi.org/10.1080/10888438.2013.774005

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine*, *15*(11), 988–994.

Fang, Y., Roscoe, R. D., & McNamara, D. S. (2021). *Predicting reading skills via stealth assessment using educational games*. Manuscript in preparation.

Fayer, S., Lacey, A., & Watson, A. (2017). STEM occupations: Past, present, and future. *Spotlight on Statistics*. https://stats.bls.gov/spotlight/2017/science-technology-engineering-and-mathematics-stem-occupations-past-present-and-future/pdf/science-technology-engineering-and-mathematics-stem-occupations-past-present-and-future.pdf.

Goldman, S. R., Braasch, J. L., Wiley, J., Graesser, A. C., & Brodowinska, K. (2012). Comprehending and learning from Internet sources: Processing patterns of better and poorer learners. *Reading Research Quarterly*, *47*(4), 356–381. https://doi.org/10.1002/RRQ.027

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371–395.

Gulikers, J. T. M., Kester, L., Kirschner, P. A., & Bastiaens, T. J. (2008). The effect of practice experience on perceptions of assessment authenticity, study approach, and learning outcomes. *Learning and Instruction*, *18*(2), 172–186. https://doi.org/10.1016/j.learninstruc.2007.02.012

Healy, A. F., Schneider, V. I., & Bourne, L. E. Jr. (2012). Empirically valid principles of training. In A. F. Healy & L. E. Bourne, (Eds.), *Training cognition: Optimizing efficiency, durability, and generalizability* (pp. 13–39). Psychology Press.

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology*, *105*(4), 1036–1049. https://doi.org/10.1037/a0032580

Jackson, G. T., & McNamara, D. S. (2017). The motivation and mastery cycle framework: Predicting long-term benefits of educational games. In Y. Baek (Ed.), *Game-based learning: Theory, strategies and performance outcomes* (pp. 97–122). Nova Science Publishers.

Jackson, G. T., Dempsey, K. B., & McNamara, D. S. (2012). Game-based practice in a reading strategy tutoring system: Showdown in iSTART-ME. In H. Reinders (Ed.), *Computer games* (pp. 115–138). Multilingual Matters.

Jackson, G. T., Snow, E. L., Varner (Allen, L. K., & McNamara, D. S. (2013). Game performance as a measure of comprehension and skill transfer. In C. Boonthum-Denecke & G. M. Youngblood (Eds.), *Proceedings of the 26th Annual Flordia Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 497–502). The AAAI Press.

Jackson, G. T., Varner, L. K., Denecke, C. B., & McNamara, D. S. (2013). The impact of individual differences on learning with an educational game and a traditional ITS. *International Journal of Learning Technology*, *8*(4), 315–336. https://doi.org/10.1504/IJLT.2013.059129

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

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). Springer.

Jacovina, M. E., Jackson, G. T., Snow, E. L., & McNamara, D. S. (2016). Timing game-based practice in a reading comprehension strategy tutor. In A. Micarelli, J. Stamper, & K. Panourgia (Eds.), *Proceedings of the 13th International Conference on Intelligent Tutoring Systems (ITS 2016)* (pp. 80–89). Springer.

James, L. T., & Casidy, R. (2018). Authentic assessment in business education: Its effects on student satisfaction and promoting behavior. *Studies in Higher Education*, *43*(3), 401–415. https://doi.org/10.1080/03075079.2016.1165659

Johnson, A. M., Guerrero, T. A., Tighe, E. L., McNamara, D. S. (2017, June). *iSTART-ALL: Confronting adult low literacy with intelligent tutoring for reading comprehension* [Paper presentation]. International Conference on Artificial Intelligence in Education (pp. 125–136). Springer.

Johnson, A. M., Jacovina, M. E., Russell, D. E., & Soto, C. M. (2016). Challenges and solutions when using technologies in the classroom. In S. A. Crossley & D. S. McNamara (Eds.), *Adaptive educational technologies for literacy instruction* (pp. 13–29). Taylor & Francis.

Johnson, A. M., Ozogul, G., & Reisslein, M. (2015). Supporting multimedia learning with visual signalling and animated pedagogical agent: Moderating effects of prior knowledge. *Journal of Computer Assisted Learning*, *31*(2), 97–115. https://doi.org/10.1111/jcal.12078

Johnson, A. M., McCarthy, K. S., Kopp, K., Perret, C. A., & McNamara, D. S. (2017). Adaptive reading and writing instruction in iSTART and W-Pal. In Z. Markov & V. Rus (Eds.), *Proceedings of the 30th Annual Florida Artificial Intelligence Research Society International Conference (FLAIRS)* (pp. 561–566). AAAI Press.

Kastberg, D., Chan, J. Y., & Murray, G. (2016). *Performance of US 15-year-old students in science, reading, and mathematics literacy in an international context: First look at PISA 2015* [NCES 2017-048]. National Center for Education Statistics.

Kendeou, P. (2015). A general inference skill. In E. J. O'Brien, A. E. Cook, & R. F. Lorch (Eds.), *Inferences during reading* (pp. 160–181). Cambridge University Press.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. https://doi.org/10.1037/0033-295x.95.2.163

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Long, D. L., Oppy, B. J., & Seely, M. R. (1994). Individual differences in the time course of inferential processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1456–1470. https://doi.org/10.1037/0278-7393.20.6.1456

MacGinitie, W. H., & MacGinitie, R. K. (1989). *Gates–MacGinitie reading tests*. Riverside.

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *The American Psychologist*, *63*(8), 760–769. https://doi.org/10.1037/0003-066X.63.8.760

Mayer, R. E. (2021). Evidence-based principles for how to design effective instructional videos. *Journal of Applied Research in Memory and Cognition*, *10*(2), 229–240. https://doi.org/10.1016/j.jarmac.2021.03.007

McCarthy, K. S., Allen, L. K., Hinze, S. R. (2020). *Predicting reading comprehension from constructed responses: Explanatory retrievals as stealth assessment* [Paper presentation]. In Proceedings of the International Conference on Artificial Intelligence in Education (pp. 197–202). Springer.

McCarthy, K. S., Watanabe, M., & McNamara, D. S. (2020). The design implementation framework: Guiding principles for the redesign of a reading comprehension intelligent tutoring system. In M. Schmidt, A. Tawfik, Y. Earnshaw, & I. Jahnke (Eds.), *Learner and User Experience Research: An introduction for the Field of Learning Design & Technology*. EdTech Books. https://edtechbooks.org/ux/9_the_design_impleme

McCarthy, K. S., Watanabe, M., Dai, J., & McNamara, D. S. (2020). Personalized learning in iSTART: Past modifications and future design. *Journal of Research on Technology in Education*, *52*(3), 301–321. https://doi.org/10.1080/15391523.2020.1716201

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*(1), 1–30. https://doi.org/10.1207/s15326950dp3801_1

McNamara, D. S. (2009). The importance of teaching reading strategies. *Perspectives on Language and Literacy*, *35*(2), 34–40.

McNamara, D. S. (2017). Self-explanation and reading strategy training (SERT) improves low-knowledge students' science course performance. *Discourse Processes*, *54*(7), 479–492. https://doi.org/10.1080/0163853X.2015.1101328

McNamara, D. S., & Healy, A. F. (1995). A procedural explanation of the generation effect: The use of an operand retrieval strategy for multiplication and addition problems. *Journal of Memory and Language*, *34*(3), 399–416. https://doi.org/10.1006/jmla.1995.1018

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247–288. https://doi.org/10.1080/01638539609544975

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 297–384). Academic Press.

McNamara, D. S., & McDaniel, M. A. (2004). Suppressing irrelevant information: Knowledge activation or inhibition? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(2), 465–482. https://doi.org/10.1037/0278-7393.30.2.465

McNamara, D. S., Jacovina, M. E., Snow, E. L., & Allen, L. K. (2015). From generating in the lab to tutoring systems in classrooms. *The*

American Journal of Psychology, *128*(2), 159–172. https://doi.org/10.5406/amerjpsyc.128.2.0159

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). Erlbaum.

McNamara, D. S., O'Reilly, T. P., Best, R. M., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research*, *34*(2), 147–171. https://doi.org/10.2190/1RU5-HDTJ-A5C8-JVWE

McNamara, D. S., Ozuru, Y., & Floyd, R. G. (2011). Comprehension challenges in the fourth grade: The roles of text cohesion, text genre, and readers' prior knowledge. *International Electronic Journal of Elementary Education*, *4*, 229–257.

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across genres and grades. In J. P. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89–116). R&L Education.

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

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, *23*(3), 187–194. https://doi.org/10.1177/01466219922031310

Oakhill, J. (1984). Inferential and memory skills in children's comprehension of stories. *British Journal of Educational Psychology*, *54*(1), 31–39. https://doi.org/10.1111/j.2044-8279.1984.tb00842.x

Oakhill, J., Cain, K., & Elbro, C. (2019). Reading comprehension and reading comprehension difficulties. In D. Kilpatrick, R. Joshi, & R. Wagner (Eds.), *Reading development and difficulties* (pp. 83–115). Springer.

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). Springer.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research*, *66*(2), 181–221. https://doi.org/10.3102/00346543066002181

Rowe, M., & McNamara, D. S. (2008). Inhibition needs no negativity: Negativity links in the construction-integration model. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1777–1782). Cognitive Science Society.

Rupley, W. H., Blair, T. R., & Nichols, W. D. (2009). Effective reading instruction for struggling readers: The role of direct/explicit teaching. *Reading & Writing Quarterly*, *25*(2–3), 125–138. https://doi.org/10.1080/10573560802683523

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences*. Cambridge University Press.

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, *74*, 14–31. https://doi.org/10.1016/j.ijhcs.2014.09.006

Shute, V. J., & Rahimi, S. (2020). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*, 1–13. https://doi.org/10.1016/j.chb.2020.106647

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. The MIT Press.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's playground. *The Journal of Educational Research*, *106*(6), 423–430. https://doi.org/10.1080/00220671.2013.832970

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117. https://doi.org/10.1016/j.chb.2016.05.047

Sithole, A., Chiyaka, E. T., McCarthy, P., Mupinga, D. M., Bucklein, B. K., & Kibirige, J. (2017). Student attraction, Persistence and retention in STEM programs: Successes and continuing challenges. *Higher Education Studies*, *7*(1), 46–59. https://doi.org/10.5539/hes.v7n1p46

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. RAND Education.

Snow, E. L., Jackson, G. T., Varner, L. K., & McNamara, D. S. (2013a). Investigating the effects of off-task personalization on system performance and attitudes within a game-based environment. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 272–275). Springer.

Snow, E. L., Jackson, G. T., Varner, L. K., & McNamara, D. S. (2013b). The impact of system interactions on motivation and performance. In *Proceedings of the 15th International Conference on Human-Computer Interaction (HCII)* (pp. 103–107). Springer.

Snow, E. L., Jacovina, M. E., Allen, L. K., Dai, J., & McNamara, D. S. (2014). Entropy: A stealth assessment of agency in learning environments. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 241–244). International Educational Data Mining Society.

Sotiriadou, P., Logan, D., Daly, A., & Guest, R. (2020). The role of authentic assessment to preserve academic integrity and promote skill development and employability. *Studies in Higher Education*, *45*(11), 2132–2148. https://doi.org/10.1080/03075079.2019.1582015

Stevens, E. A., Park, S., & Vaughn, S. (2019). A review of summarizing and main idea interventions for struggling readers in grades 3 through 12: 1978–2016. *Remedial and Special Education*, *40*(3), 131–149. https://doi.org/10.1177/0741932517749940

Tsai, C. Y., Lin, H. S., & Liu, S. C. (2020). The effect of pedagogical GAME model on students' PISA scientific competencies. *Journal of Computer Assisted Learning*, *36*(3), 359–369. https://doi.org/10.1111/jcal.12406

Van Den Broek, P., Rapp, D. N., & Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Processes*, *39*(2–3), 299–316. https://doi.org/10.1080/0163853X.2005.9651685

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

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568–2572. https://doi.org/10.1016/j.chb.2013.06.033

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, *227*, 483–493.

Wang, F., Li, W., Mayer, R. E., & Liu, H. (2018). Animated pedagogical agents as aids in multimedia learning: Effects on eye-fixations during learning and learning outcomes. *Journal of Educational Psychology*, *110*(2), 250–268. https://doi.org/10.1037/edu0000221

Wolfe, M. B., & Goldman, S. R. (2005). Relations between adolescents' text processing and reasoning. *Cognition and Instruction*, *23*(4), 467–502. https://doi.org/10.1207/s1532690xci2304_2

Xue, Y., Larson, R. C. (2015). *Stem crisis or stem surplus? yes and yes: Monthly labor review*. U.S. Bureau of Labor Statistics. Retrieved February 9, 2022, from https://www.bls.gov/opub/mlr/2015/article/stem-crisis-or-stem-surplus-yes-and-yes.htm

About the authors

Danielle S. McNamara develops educational technologies (iSTART, iSTART-ME, Coh-Metrix, Writing-Pal) and conducts research to better understand cognitive processes of comprehension, learning, text coherence, and individual differences. She has solidified herself as a one of the world's premier experts in cognitive psychology, publishing hundreds of scholarly works.

Tracy Arner is a postdoctoral research scholar at Arizona State University. Her research interests include the development and use of

instructional technologies such as multimedia instruction, intelligent tutoring systems, and artificial intelligence to improve learning outcomes for struggling students.

Reese Butterfuss is a postdoctoral research scholar at Arizona State University. His research focuses on the cognitive processes that underlie learning from texts, as well as improving students' literacy skills using technology-based literacy instruction.

Ying Fang is an associate professor of Faculty of Artificial Intelligence in Education at Central China Normal University. Her research interests include artificial intelligence in educational systems, the development, implementation, and assessment of intelligent tutoring systems, and promoting learning in electronic learning environments.

Micah Watanabe is a PhD candidate in Cognitive Psychology at Arizona State University. His research is primarily on the role of educational interventions and intelligent tutoring systems in promoting conceptual change. He also studies the structure and quality of students' prior knowledge and how that affects their learning.

Natalie Newton is a Research Specialist at Arizona State University. She leads training for expert raters to score constructed responses and contributes to other research projects through data analysis, writing, and conceptualization. Her research interests in the lab center around comprehension strategy use and how instructional prompts impact text understanding.

Katie S. McCarthy is an Assistant Professor of Educational Psychology in the Department of Learning Sciences at Georgia State University. Her research explores the higher-order processes involved in reading comprehension and how they vary across disciplines and readers and how in-person interventions and educational technology can support learning from text.

Laura K. Allen is an Assistant Professor of Educational Psychology at University of Minnesota. The primary aim of her research is to examine how individuals learn and communicate with text and to apply those insights to educational practice through the development of interventions and educational technologies.

Rod D. Roscoe is an Associate Professor of Human Systems Engineering in the Ira A. Fulton Schools of Engineering at Arizona State University. His work combines insights from learning science, cognitive science, design science, and equity science to implement effective educational technologies that are inclusive for all learners.