computers

Article

Stealth Literacy Assessments via Educational Games

Ying Fang [1], Tong Li [2], Linh Huynh [3], Katerina Christhilf [3], Rod D. Roscoe [4] and Danielle S. McNamara [3,*]

[1] Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China; fangying@ccnu.edu.cn
[2] School of Journalism and Strategic Communication, Ball State University, Muncie, IN 47306, USA
[3] Department of Psychology, Arizona State University, Tempe, AZ 85281, USA
[4] Human Systems Engineering, Arizona State University, Mesa, AZ 85212, USA
* Correspondence: dsmcnamara1@gmail.com

Abstract: Literacy assessment is essential for effective literacy instruction and training. However, traditional paper-based literacy assessments are typically decontextualized and may cause stress and anxiety for test takers. In contrast, serious games and game environments allow for the assessment of literacy in more authentic and engaging ways, which has some potential to increase the assessment's validity and reliability. The primary objective of this study is to examine the feasibility of a novel approach for stealthily assessing literacy skills using games in an intelligent tutoring system (ITS) designed for reading comprehension strategy training. We investigated the degree to which learners' game performance and enjoyment predicted their scores on standardized reading tests. Amazon Mechanical Turk participants (n = 211) played three games in iSTART and self-reported their level of game enjoyment after each game. Participants also completed the Gates–MacGinitie Reading Test (GMRT), which includes vocabulary knowledge and reading comprehension measures. The results indicated that participants' performance in each game as well as the combined performance across all three games predicted their literacy skills. However, the relations between game enjoyment and literacy skills varied across games. These findings suggest the potential of leveraging serious games to assess students' literacy skills and improve the adaptivity of game-based learning environments.

Keywords: literacy; assessment; reading comprehension; educational games; intelligent tutoring system

Citation: Fang, Y.; Li, T.; Huynh, L.; Christhilf, K.; Roscoe, R.D.; McNamara, D.S. Stealth Literacy Assessments via Educational Games. *Computers* **2023**, *12*, 130. https://doi.org/10.3390/computers12070130

Academic Editors: Carlos Vaz de Carvalho, Hariklia Tsalapatas and Ricardo Baptista

Received: 23 May 2023
Revised: 18 June 2023
Accepted: 20 June 2023
Published: 25 June 2023

1. Introduction

Literacy can be broadly defined as "the ability to understand, evaluate, use, and engage with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential" [1]. Literacy skills are critical to academic success and in life; however, large-scale reading assessment data reveal that many students and adults struggle with reading comprehension. The most recent National Assessment of Educational Progress reported that 27% of 8th grade students in the United States performed below the basic levels of reading comprehension and 66% did not reach proficient levels. Similarly, 30% and 63% of 12th graders did not reach basic and proficiency reading levels, respectively [2]. As might be expected, deficits in reading skills often continue into adulthood. According to the 2017 Program for the International Assessment of Adult Competencies assessment, 19% of U.S. adults aged 16 or older performed at or below the lowest literacy level [3].

Literacy assessments are a key component of any effort to improve students' literacy or remediate potential gaps in education. A good understanding of learners' current skills reveals the types and amounts of instruction they will need to grow. However, traditional literacy assessments (e.g., standardized tests) typically require a significant amount of time to administer, score, and interpret. Additionally, these tests usually occur before or after learning, making it difficult to provide timely feedback to guide teaching and learning [4–6]. Moreover, traditional assessments may cause stress and test anxiety, which may in turn negatively impact students' test-taking experiences [5,7], and change how students respond to the assessments [8].

In contrast to traditional literacy assessments, stealth assessment offers an innovative approach by implementing literacy assessments in computer-based learning environments. These assessments take place *during* learning activities, instead of summative or "checkpoint" assessments. In addition, the assessments are based on students' natural behaviors and performance rather than being presented as "tests." As such, stealth literacy assessments can evaluate student reading skills unobtrusively and dynamically, and provide timely feedback throughout the learning process. In this innovative context, serious games have strong potential to be more motivating and enjoyable than traditional reading assessments. Thus, in this study, we investigated the feasibility of game-based stealth assessment to predict literacy skills, specifically reading comprehension and vocabulary knowledge.

1.1. Stealth Assessment within Games

Stealth assessment refers to performance-based assessments that are seamlessly embedded in gaming environments without the awareness of students who are being evaluated [4,9]. Stealth assessment was initially proposed and explored to assess higher-order competencies such as persistence, creativity, self-efficacy, openness, and teamwork, primarily because these competencies substantially impact student academic achievement, but also because traditional methods of assessment often neglect these abilities [10,11]. As such, stealth assessments that analyze how students use knowledge and skills during gameplay have been embedded in serious games to unobtrusively assess those competencies [12–15]. Game environments may make assessment less salient or less visible, and thus students feel that they are merely "playing" rather than "being tested".

In stealth assessment, traditional test items are replaced by authentic, real-world scenarios or game tasks. Since stealth assessment items can be contextualized and potentially connected to the real world, students' skills, behaviors, and competencies may be more validly demonstrated through these game activities than in traditional assessments [16,17]. Within stealth assessments, students generate rich sequences of performance data (e.g., choices, actions, and errors) when they perform the tasks, which can serve as evidence for knowledge and skills assessment. When students are assessed without the feeling of being tested, it can reduce their stress and anxiety, which can in turn increase the reliability of the assessment [18,19].

Serious games in education are designed to enhance students' learning experience by providing a more fun way to acquire knowledge [20,21], which could be ideal for stealth assessments because they further separate students' experiences of play and enjoyment from experiences of testing and measurement. In a well-designed game, students are immersed in game scenarios and motivated to proceed through the challenges and meet learning goals, which might not feel like a learning or testing experience at all [22]. For example, Physics Playground is a game that emphasizes 2-D physics simulations. The game implements stealth assessments to evaluate students' physics knowledge, persistence, and creativity [15,23]. When students interact with the game, they produce a dense stream of performance data, which is recorded by the system in a log file and analyzed using Bayesian networks to infer students' knowledge and skills. The system then provides ongoing automated feedback to teachers and/or students, based on the assessment, to support student learning. Another example of stealth assessment is a game-based learning environment named ENGAGE that was designed to promote computational thinking skills. Students' behavioral data were collected during their gameplay and then analyzed using machine learning methods to infer their problem-solving skills and computational knowledge [13,24].

1.2. Stealth Reading Assessment via Games

Prior studies have explored stealth assessment via games to assess students' higher-order skills and competencies, such as problem solving, creativity, and persistence, along with scientific knowledge [14,17,23,25]. Only a few studies have investigated using stealth assessment to assess *literacy* [26–28]. These studies primarily leveraged natural language processing (NLP) techniques to extract linguistic properties of constructed responses (e.g., essays and explanations) to make predictions about students' reading skills. For example, Allen and McNamara analyzed the lexical properties of students' essays to predict students' vocabulary test scores. Two lexical indices associated with the use of sophisticated and academic word use accounted for 44% of the variance in vocabulary knowledge scores [26]. Fang et al. predicted students' comprehension test scores using the linguistic properties of the self-explanations generated during their practice in a game-based learning environment. Five linguistic features accounted for around 20% of the variance in comprehension test scores across datasets. The studies collectively demonstrate linguistic features at multiple levels of language (e.g., lexical, syntactic, and semantic) have strong potential to serve as proxies of reading skills, supporting the feasibility of stealth reading assessment using NLP [27].

However, the use of NLP for reading assessment is not without challenges. These methods rely on machine learning algorithms, which require a large amount of data to train and test to improve the analysis accuracy [29]. Additionally, some NLP methods rely on extensive computational resources to support the complex calculations, which might be difficult to adopt by development teams without those resources [29]. Most importantly, NLP is language-specific, rendering it challenging to generalize algorithms across languages.

An alternative to NLP is the analysis of students' performance data from games implementing multiple-choice questions. Multiple-choice questions provide a shortlist of answers for students to choose from, which do not require complex data analysis to evaluate students' answers. For example, Fang et al. investigated the association between students' reading skills and their performance in a *single* vocabulary game (i.e., Vocab Flash). The analysis was based on students' performance data from the game implementing what are essentially multiple-choice questions in disguise. The results of the study supported the value of using a simple vocabulary game to assess reading comprehension [30].

The current study examines both vocabulary and main idea games. In the following section, we introduce how literacy skills are reflected by the ability to identify word meaning and text main ideas, providing the theoretical grounds for leveraging vocabulary and main idea games to assess reading skills.

1.3. Assessing Reading Skills through Vocabulary and Main Idea Games

Reading skills are at least partially reflected by students' vocabulary knowledge and ability to identify main ideas of passages. Readers must be able to process the basic elements of the text, including the individual words and the syntax, to understand and gain meaning from texts. From those elements, the reader can construct an understanding of the meanings behind phrases and sentences. Readers with more vocabulary knowledge tend to have better reading comprehension skills [31,32]. When a reader is unfamiliar with certain keywords in a text, this can slow down or fully impair the processing of key points in the text. Although in some cases the meaning of words can be understood via contextual cues, comprehension becomes more challenging for readers with lower vocabulary knowledge [33]. This effect is exacerbated when the reader also has insufficient prior knowledge about the topic through which to build context [34]. Hence, when assessing texts for readability, educators and researchers have found that texts containing many low-frequency and sophisticated words make them more difficult to read [35]. In second-language learners, vocabulary is one of the most critical factors in determining how well students can comprehend texts [36,37].

Comprehending text requires not only knowledge of individual word meanings, but also the skills required to deduce relations between ideas and, in turn, the main ideas [38]. Identifying topic sentences is a comprehension strategy that requires students to recognize the main ideas within a text while dismissing information that is irrelevant or redundant [39]. Being able to distinguish main versus supporting ideas can foster deep comprehension because it encourages students to attend to the higher-level meaning and the global organization of information across texts [40]. Consequently, Bransford et al. suggested that identifying main ideas within a text can lead to enhanced comprehension and retention of the text content [40]. According to Wade-Stein and Kintsch [41], this task not only promotes students' construction of factual knowledge, but also of conceptual knowledge, as the process of identifying main ideas within a text reinforces students' memory representations of its content.

Deducing the gist of a text can be challenging [42,43]. Low-knowledge readers can find it hard to differentiate between the main arguments and supporting ideas of a text [44]. Likewise, Wigent found that students with reading difficulties focused more on details rather than identifying main ideas, subsequently recognizing and recalling fewer topic sentences compared to average readers [45]. By contrast, strategic and skilled readers are more likely to grasp the main ideas from text compared to less skilled readers [46–49].

Students' reading comprehension skills, vocabulary knowledge, and ability to recognize main ideas are closely associated. Therefore, this study employed three distinct games that emphasized vocabulary and main idea identification. First, Vocab Flash is a game that requires players to select appropriate synonyms for words. It is an adaptive game designed to measure vocabulary for variously skilled participants, making it ideal for stealth assessment of vocabulary. Second, Adventurer's Loot is a game that asks players to read a text and select the main ideas. Participants must be careful to select only the main ideas, and not any extraneous details. Finally, Dungeon Escape asks players to read a passage, and imagine they are about to write a summary of the passage. They must select the best topic sentence for the summary. To pick the best topic sentence, participants must locate and integrate the main ideas of the passage. Such integration may require further reading comprehension skill than Adventurer's Loot, which is why two main idea games are included.

1.4. Current Study

The goal of the current study is to investigate the feasibility of using games for stealth assessment of reading skills. Specifically, this study examines the predictive value of *three* distinct games, namely one game that targets vocabulary knowledge and two games that target main idea identification. We not only examine students' performance in each game individually, but also explore the value of combining performance data across all three games. The goal is to assess the extent to which students' performance in the three games is indicative of their reading skills as measured by standardized reading tests.

In addition to game performance, we consider students' subjective game experience. Based on the literature regarding serious games [50–52], we expect that most students will have overall positive attitudes toward playing the games. However, it is unknown whether students' game enjoyment will influence the validity of the stealth assessment. To that end, we address the following research questions in this study:

1. To what extent does students' performance in the three games predict their reading skills (i.e., vocabulary knowledge and reading comprehension)?
2. Does students' enjoyment of the games moderate the relations between game performance and reading skills?

2. Methods

2.1. Experimental Environment: iSTART

The Interactive Strategy Training for Active Reading and Thinking (iSTART) is a game-based intelligent tutoring system (ITS) designed to help students improve reading skills through adaptive instruction and training. iSTART was developed based upon Self-Explanation Reading Training (SERT) [53], a successful classroom intervention that taught students to explain the meaning of texts while reading (i.e., self-explain) through the use of comprehension strategies (i.e., comprehension monitoring, paraphrasing, predicting, bridging, and elaborating). The current version of iSTART includes three training modules focusing on self-explanation, summarization, and question asking [54].

iSTART learning materials consist of video lessons and two types of practice: regular and game-based practice. Video lessons provide students with information about comprehension strategies and prepare them for the practice. During the regular practice, students complete given tasks, and the system provides immediate feedback on students' performance. For example, when students generate a self-explanation on a target sentence, the NLP algorithms implemented in iSTART automatically analyze the self-explanation and provide real-time feedback. The feedback includes a holistic score on a scale of 0 ("poor") to 3 ("great") and actionable feedback to help students improve the self-explanation when the score is below a certain threshold [55]. Studies that investigate the effectiveness of iSTART indicate that iSTART facilitates both comprehension strategy learning and comprehension skills [53,54,56].

2.1.1. iSTART Games

iSTART implements two forms of game-based practice to increase learners' motivation and engagement: generative and identification games [54,56]. In generative games, students are asked to construct verbal responses such as self-explanations. NLP-based algorithms assess these constructed responses to determine the quality and/or the use of specific strategies. In contrast, identification games ask students to review short example stimuli or prompts, and then to choose one or more responses that correctly identify strategy use or follow from the prompts. For example, students might read an example self-explanation and then indicate whether the excerpt demonstrates "paraphrasing" or "elaborating." In a vocabulary game, students may be given a prompt term and then must choose a correct synonym from several choices. Importantly, alternatives typically comprise carefully generated foils, such that incorrect answers are diagnostic of student misunderstanding.

iSTART games also include narrative scenarios and other challenges to further motivate reading strategy practice. For instance, students are rewarded with "iBucks" during gameplay, which can be "spent" to unlock additional game backgrounds or to customize personal avatar characters [54]. In addition, students receive immediate feedback during or after gameplay to support their self-monitoring and engagement [52]. For example, in Showdown, students compete against a computer-controlled player to explain target sentences in given texts. At the end of each round, the system evaluates students' answers and informs them of their performance ("poor", "fair", "good", or "great"). Meanwhile, the performance scores of students are compared with the computer-controlled player to determine who wins the round (see Figure 1).

2.1.2. Adaptivity Facilitated by Assessments in iSTART

iSTART implements both inner-loop and outer-loop adaptivity to customize instruction to individual students. Inner-loop adaptivity refers to the immediate feedback students are given when they complete an individual task, and outer-loop adaptivity refers to the selection of subsequent tasks based on students' past performance [57].

Regarding inner-loop adaptivity, the generative games utilize NLP (e.g., LSA) and machine learning algorithms to assess constructed responses, and then provide holistic scores and actionable, individualized feedback [55,56]. Within identification games, the assessment of the answers matches students' selection with predetermined answers. The system then provides timely feedback including response accuracy, explanations of why the responses are correct or incorrect, and game performance scores.

To further promote skill acquisition, iSTART complements the inner-loop with outer-loop adaptations, which select practice texts based on the student model and the instruction model. An ITS typically employs three elements to assess students and select appropriate tasks: the domain model, the student model, and the instructional model [57–59]. The domain model represents ideal expert knowledge and may also address common student misconceptions. The domain model is usually created using detailed analyses of the knowledge elicited from subject matter experts. The student model represents students' current understanding of the subject matter, and it is constructed by examining student task performance in comparison to the domain model. Finally, the instructional model represents the instructional strategies. It is used to select instructional content or tasks based on inferences about student knowledge and skills. iSTART creates student models using students' self-explanation scores and scores on multiple-choice measures. The instructional model then determines the features of each presented task (i.e., text difficulty and scaffolds to support comprehension) using the evolving student model. For example, subsequent texts become more difficult if students' self-explanation quality on prior texts is higher. Conversely, when students' self-explanation quality is lower, the subsequent texts become easier [54].

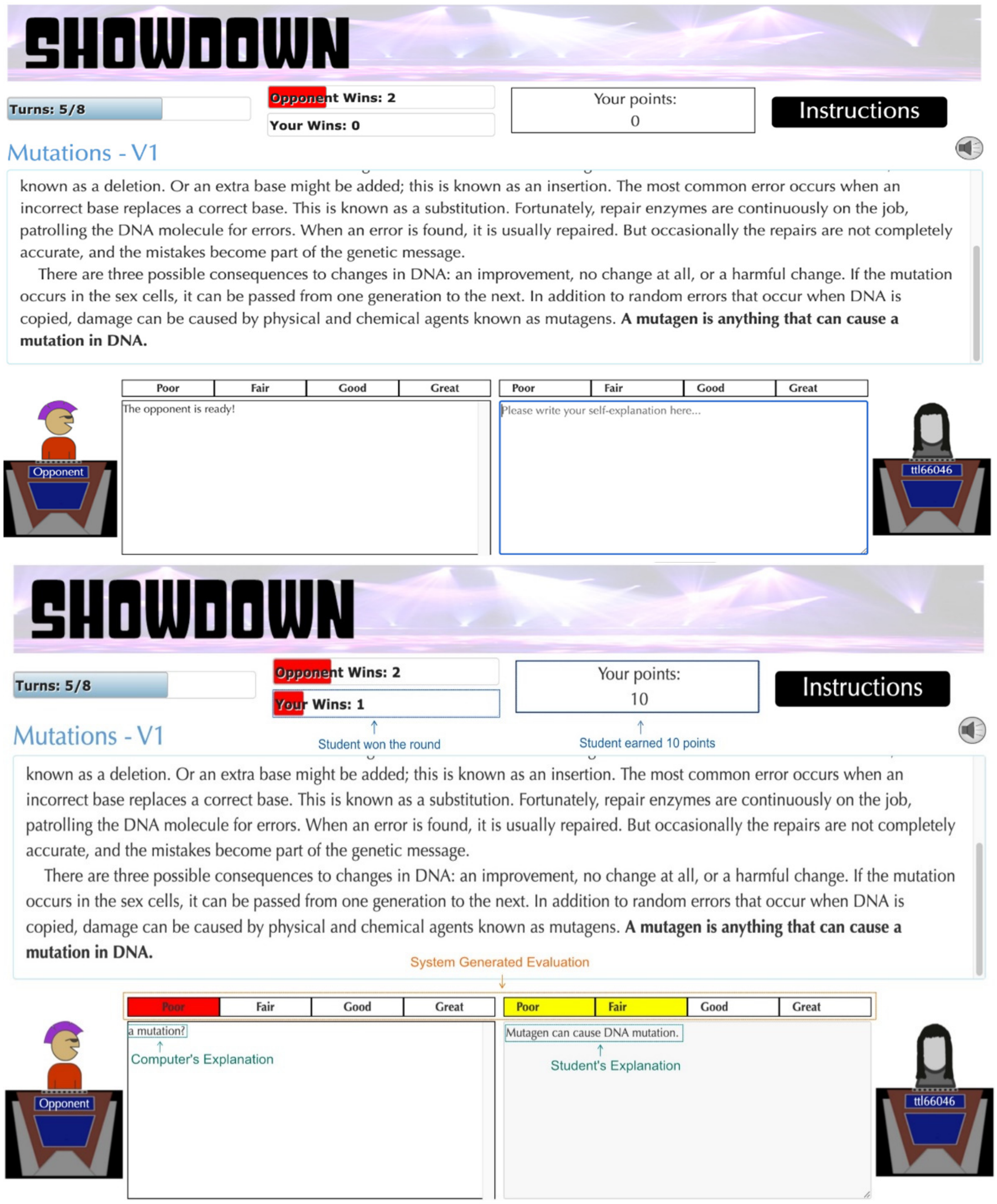

Figure 1. Showdown is a practice game within iSTART. The first image (**top**) shows the game interface where the human player and computer-controlled player are about to start the competition. The second image (**bottom**) shows both players' explanations and the system-generated evaluations of the explanations.

Current adaptation facilitated by assessments in iSTART is at the micro-level. Specifically, the assessments are task-specific and may not transfer to the games implementing different types of tasks. For example, students' scores on self-explanation games may not reflect their performance in question-asking games. As such, students' self-explanation scores are not leveraged to guide the adaptivity (e.g., learning material selection) in the question-asking module. We anticipate that task-general assessments, such as reading skill assessments, have strong potential to supplement current assessments to guide the macro-level adaptation across modules.

2.2. Participants

Participants were 246 adults recruited on Amazon Mechanical Turk (an online platform). Thirty-five participants were excluded from the study due to failing an attention check, which resulted in the final sample of 211 participants (98 female, 113 male). In the final sample, 77.2% of participants identified as Caucasian, 10.9% as African American, 6.6% as Hispanic, 4.3% as Asian, and 1.0% as another race/ethnicity. Participants were 37.2 years old, on average, with a range of 17 to 68. Most participants (81.5%) reported holding a Bachelor's or advanced degree.

2.3. Procedure, Materials, and Measures

Participants first responded to a demographic questionnaire, and then played three iSTART games: Vocab Flash, Dungeon Escape, and Adventurer's Loot. Game order was counterbalanced. After every game, participants completed a brief questionnaire regarding their enjoyment. In the final step of the study, participants completed the Gates–MacGinitie Reading Test (GMRT), which included a vocabulary and a comprehension subtests.

Gates–MacGinitie Reading Test (GMRT). Participants' reading skills were measured by the Gates–MacGinitie Reading Test (GMRT) level 10/12 form S. The GMRT is an established and reliable measure of reading comprehension (α = 0.85–0.92) [60], which comprises both vocabulary and comprehension subtests. The vocabulary subtest (10 min) includes 45 multiple-choice questions that ask participants to choose the correct definition of target words in the given sentences. The comprehension subtest (20 min) consists of a series of textual passages with two to six multiple-choice comprehension questions per passage. There are a total of 48 questions.

Vocabulary and reading comprehension skills were operationally defined as the total number of correct answers on the vocabulary and reading comprehension GMRT subtests, respectively.

Vocab Flash. In this game, students read a target word and must choose a synonym out of four alternatives (i.e., one correct choice and three incorrect foils). Students are allotted 5 min to respond to as many terms as possible. For each target word, students are only allowed one attempt to select the answer. After students submit the answer, they receive feedback that (a) indicates whether their answer is correct and (b) clearly highlights the correct response. One key feature of the game is its adaptivity. The target words are classified into nine different levels of difficulty based on their frequency rating in Corpus of Contemporary American English (COCA) [61]. Uncommon words are typically more challenging. The game begins with the easiest words and progresses to higher levels of difficulty as students answer correctly. However, students can also return to easier levels after repeated errors. As in computer-adaptive testing [62], students can fluctuate between levels of difficulty, but more skilled students will generally encounter more difficult items. Game performance in Vocab Flash was measured by the proportion of correctly answered questions.

Dungeon Escape. Dungeon Escape is an iSTART game in which students are knights trapped in a dungeon. The way to escape it is to earn points by selecting topic sentences of given texts. Each student must complete six texts that are randomly selected from the game's science text pool. Four alternative sentences (i.e., one correct answer and three incorrect foils) are provided for each text. Students are allowed multiple attempts for each question, and they proceed to the next text by selecting the correct answer. Performance in Dungeon Escape was measured by the proportion of correct answers only in students' first attempts, because students may potentially game the system (e.g., try all of the answers sequentially).

Adventurer's Loot. Adventurer's Loot is an iSTART game in which students are asked to discover the hidden treasures on a map by selecting the main ideas of given texts. There are eight sites on the map, and each site corresponds to a specific text. Students can select a site from the map to explore and work on the corresponding text. Students are allowed multiple attempts on a text. The only way to proceed to the next text is by answering the question correctly, namely, selecting all the main ideas. Importantly, the number of correct answers (i.e., main ideas) in this game varies between texts. For the texts with multiple correct answers, the incorrect answers may be missing main ideas or selecting distractors. To be sensitive to different error types and potential user attempts to game the system, d prime was used as the performance measure of Adventure's Loot. It was based on (1) the proportion of correctly identified main ideas in the first attempt, which was computed using the number of correctly selected answers divided by the number of correct answers, and (2) the proportion of incorrectly selected distractors in the first attempt, which was computed using the number of incorrectly selected answers (i.e., selected distractors) divided by the number of distractors. D prime was calculated using the z score of (1), subtracting the z score of (2).

Game Survey. After each game, participants responded to six items pertaining to their subjective game enjoyment. These questions were derived from measures implemented in prior studies: (1) This game was fun to play; (2) This game was frustrating; (3) I enjoyed playing this game; (4) This game was boring; (5) The tasks in this game were easy; and (6) I would play this game again. Participants rated their agreement with these statements on a 6-point Likert scale ranging from "1" (strongly disagree) to "6" (strongly agree). Student enjoyment of the game was operationalized as the average score of the six items.

2.4. Statistical Analyses

Internal consistency between survey items (i.e., reliability) was measured using Cronbach's alpha calculated with the following formula:

$$\alpha = \frac{N \times \bar{c}}{\bar{v} + (N-1) \times \bar{c}}$$

where N = number of items, $\bar{c}$ = mean covariance between items, and $\bar{v}$ = mean item variance. Two items "This game was frustrating" and "This game was boring" were reverse coded before the calculation such that all of the items indicated positive attitudes toward the games.

Hierarchical linear regressions were conducted to determine whether student game performance and enjoyment predict reading test performance. More specifically, game performance scores and enjoyment scores for each game (Vocab Flash, Dungeon Escape, and Adventurer's Loot) were used to predict vocabulary test scores and comprehension test scores.

Finally, hierarchical linear regressions were conducted to examine whether participants' performance and enjoyment *combined* across all three games were better predictors of reading skills than performance and enjoyment of each *individual* game.

3. Results

3.1. Survey Item Internal Consistency

Cronbach's alpha of the six survey items for Vocab Flash, Dungeon Escape, and Adventurer's Loot were 0.68, 0.73, and 0.82, respectively. A general accepted rule is that alpha of 0.6–0.7 indicates an acceptable level of reliability, and 0.8 or greater a good level [63]. Therefore, the scores indicated acceptable to good internal consistency between the survey items.

3.2. Descriptive Statistics of Predictor and Predicted Variables

Table 1 provides descriptive statistics of participants' performance on the vocabulary and comprehension subtests, as well as their performance and enjoyment scores for the three games. Reading tests performance scores were calculated using the proportion of correct answers in the subtests. Game performance scores were calculated using the proportion of correct answers or the proportion of correct and incorrect answers, depending on the game. Game enjoyment scores were calculated using the sum of participants' ratings on the game enjoyment survey. As is shown in Table 1, participants' vocabulary and comprehension test scores were strongly and positively correlated (r = 0.76). The correlation between game performance scores and reading test scores were different for each game. Students tended to enjoy playing the games, particularly Vocab Flash (M = 4.34). However, the strength and direction of the correlations between game enjoyment and participants' reading test scores varied between games.

Table 1. Descriptive statistics and correlations between predictor and predicted variables.

Measure	*M*	*SD*	Vocabulary	Comprehension	VF Correct	DE Correct	AL Correct	AL Incorrect	VF Enjoyment	DE Enjoyment
Vocabulary	0.43	0.29								
Comprehension	0.36	0.22	0.76 **							
VF Correct	0.48	0.24	0.76 **	0.60 **						
DE Correct	0.41	0.23	0.46 **	0.50 **	0.38 **					
AL Correct	0.61	0.25	0.09	0.08	0.11	0.16 *				
AL Incorrect	0.55	0.26	−0.27 **	−0.26 **	−0.21 *	−0.07	0.67 **			
VF Enjoyment	4.34	0.84	0.13	0.16 *	0.24 **	0.07	0.03	−0.10		
DE Enjoyment	4.11	0.99	−0.26 **	−0.23 *	−0.13	0.00	−0.04	0.07	0.45 **	
AL Enjoyment	3.75	1.16	−0.60 **	−0.52 **	−0.41 **	−0.38 **	−0.10	0.12	0.31 **	0.63 **

Note. M = mean, SD = standard deviation, VF = Vocab Flash, DE = Dungeon Escape, AL = Adventurer's Loot, ** $p < 0.01$, * $p < 0.05$.

3.3. Predicting Vocabulary Knowledge with Individual Game Performance

Using hierarchical linear regression analyses, we explored whether vocabulary test scores could be predicted based on game performance and enjoyment for each game. Game performance measures were entered as predictors in Model 1, and then both game performance and enjoyment were entered as predictors in Model 2. More specifically, performance measures refer to the proportion of correct answers in Vocab Flash, proportion of correct answers for the first attempts in Dungeon Escape, and d prime scores for the first attempts in Adventurer's Loot (the calculation of d prime scores is introduced in Section 2.3). Enjoyment refers to the sum of participants' self-reported scores on the survey items. As is shown in Table 2, participants' performance in game Vocab Flash was a strong predictor and explained 57% of the variance in their vocabulary test scores. Their enjoyment of the game did not account for additional variance. For Dungeon Escape and Adventurer's Loot, participants' performance scores were again significant predictors of their vocabulary test scores. Their performance scores accounted for 20% of the variance in both games. The additional variance explained by game enjoyment was higher in Adventurer's Loot (24%) than in Dungeon Escape (6%).

3.4. Predicting Comprehension Test Scores with Individual Game Performance

As with vocabulary, hierarchical linear regression analyses sought to predict comprehension test scores based on game performance and enjoyment for Vocab Flash, Dungeon Escape, and Adventurer's Loot. In Model 1, game performance was entered as the sole predictor of comprehension test scores. Specifically, performance was measured by participants' proportion of correct answers in Vocab Flash, proportion of correct answers for the first attempts in Dungeon Escape, and d prime scores for the first attempts in Adventurer's Loot. In Model 2, both performance and enjoyment were entered as predictors of comprehension. For Vocab Flash, participants' performance scores explained 36% of the variance in comprehension test scores, with enjoyment adding no extra variance. For Dungeon Escape and Adventurer's Loot (see Table 3), participants' performance scores were significant predictors, which accounted for 25% and 18% of the variance in the comprehension test scores, respectively. The additional variance for which enjoyment accounted was 2% in Dungeon Escape and 18% in Adventurer's Loot.

Table 2. Regression analysis predicting vocabulary test scores with individual game performance and enjoyment.

Variable	Standardized Coefficient	t	R^2	R^2 Change
Vocab Flash				
Model 1			0.57	0.57 ***
Performance	0.76	16.68 ***		
Model 2			0.57	0.00
Performance	0.77	16.46 ***		
Enjoyment	−0.05	1.07		
Dungeon Escape				
Model 1			0.20	0.20 ***
Performance	0.45	6.95 ***		
Model 2			0.26	0.06 ***
Performance	0.45	7.20 ***		
Enjoyment	−0.25	−4.05 ***		
Adventurer's Loot				
Model 1			0.20	0.20 ***
Performance	0.44	7.18 ***		
Model 2			0.44	0.24 ***
Performance	0.31	5.72 ***		
Enjoyment	−0.51	−9.51 ***		

Note. *** $p < 0.001$.

Table 3. Regression analysis predicting comprehension test scores with individual game performance and enjoyment.

Variable	Standardized Coefficient	t	R^2	R^2 Change
Vocab Flash				
Model 1			0.36	0.36 ***
Performance	0.60	8.86 ***		
Model 2			0.36	0.00
Performance	0.60	8.49 ***		
Enjoyment	−0.01	0.16		
Dungeon Escape				
Model 1			0.25	0.25 ***
Performance	0.60	6.60 ***		
Model 2			0.27	0.02 ***
Performance	0.48	6.36 ***		
Enjoyment	−0.17	−2.31 *		
Adventurer's Loot				
Model 1			0.18	0.18 ***
Performance	0.43	5.69 ***		
Model 2			0.36	0.18 ***
Performance	0.31	4.48 ***		
Enjoyment	−0.44	−6.27 ***		

Note. *** $p < 0.001$, * $p < 0.05$.

3.5. Predicting Vocabulary Knowledge from Performance Combined across Games

In addition to the analyses examining each game separately, a hierarchical linear regression was conducted to predict vocabulary test scores based on their performance and enjoyment across all three games combined. The performance measures in Vocab

Flash, Dungeon Escape, and Adventurer's Loot were predictors of vocabulary test scores in Model 1. More specifically, the predictors were participants' proportion of correct scores in Vocab Flash, proportion of correct scores of the first attempts in Dungeon Escape, and d prime scores of the first attempts in Adventurer's Loot. Participants' performance in all three games were significant predictors of their vocabulary test scores and explained 65% of the variance. The explained variance was higher than that explained by performance measures in any individual game. Model 2 predicted vocabulary test scores with both performance and enjoyment scores in the three games. Game enjoyment scores added 9% of explained variance in vocabulary test scores (see Table 4).

Table 4. Regression analysis predicting vocabulary test scores with combined game performance.

Variable	Standardized Coefficient	t	R^2	R^2 Change
Model 1			0.65	0.65 ***
Performance (VF)	0.65	12.84 ***		
Performance(DE)	0.17	3.51 **		
D-prime (AL)	0.16	3.18 **		
Model 2			0.74	0.09 *
Performance(VF)	0.51	10.54 ***		
Performance(DE)	0.08	1.66		
Performance (AL)	0.11	2.55 *		
Enjoyment (VF)	0.11	2.33 *		
Enjoyment (DE)	0.03	0.50		
Enjoyment (AL)	−0.40	6.53 ***		

Note. VF = Vocab Flash, DE = Dungeon Escape, AL = Adventurer's Loot, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

3.6. Predicting Comprehension Test Scores from Performance Combined across Games

A second hierarchical linear regression sought to predict comprehension test scores based on participants' game performance and enjoyment across all three games. Model 1 only included game performance measures in the three games as predictors. Results indicated that the performance scores of all three games were significant predictors of comprehension test scores and they accounted for 49% of the variance. Model 2 included both game performance and enjoyment measures in the three games as predicting variables. The significant predictors of comprehension test scores were participants' performances in Vocab Flash and Dungeon Escape and their enjoyment of Adventurer's Loot. The additional variance explained by game enjoyment beyond the performance scores was 6% (see Table 5).

Table 5. Regression analysis predicting comprehension test scores with combined game performance and enjoyment.

Variable	Standardized Coefficient	t	R^2	R^2 Change
Model 1			0.49	0.49 ***
Performance (VF)	0.43	5.86 ***		
Performance (DE)	0.30	4.35 ***		
Performance (AL)	0.18	2.59 *		
Model 2			0.55	0.06
Performance (VF)	0.30	3.66 ***		
Performance(DE)	0.20	2.77 **		
Performance (AL)	0.13	1.90		
Enjoyment (VF)	0.14	1.77		
Enjoyment (DE)	0.01	0.10		
Enjoyment (AL)	−0.35	3.24 **		

Note. VF = Vocab Flash, DE = Dungeon Escape, AL = Adventurer's Loot, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4. Discussion

In the current study, we investigated the feasibility of game-based stealth literacy assessment using games from the iSTART ITS. Specifically, we explored to what degree learners' game performance and enjoyment in three games (i.e., Vocab Flash, Dungeon Escape, and Adventurer's Loot) were able to predict their vocabulary knowledge and reading comprehension skills. In addition, we examined whether the associations between reading skills and game performance were moderated by participants' enjoyment of the games.

Our results suggest that game performance was predictive of participants' reading skills. Specifically, performance in Vocab Flash, Dungeon Escape, and Adventurer's Loot accounted for respectively 57%, 20%, and 20% of the variance of participants' vocabulary knowledge. The explained variance increased to 65% when the combined performance of all three games was used as a predictor. Performance in Vocab Flash, Dungeon Escape, and Adventurer's Loot explained respectively 36%, 25%, and 18% of the variance of participants' comprehension. The explained variance increased to 49% when using the combined performance across all three games as a predictor. These findings demonstrate that student performance in relatively well-designed reading games can provide valid measures of reading skills. As such, reading games may be a viable alternative to standardized reading tests, which can render the testing experience more enjoyable, motivating, and engaging [5,52,64]. Another benefit of using games for assessments is that students can be assessed during gameplay, without being interrupted or feeling "tested". This stealthy approach may reduce test anxiety, and in turn increase the reliability of the assessment [18,19,65].

One approach to assessing reading skills in prior research has been in the context of constructed responses wherein students generate self-explanations or essays. The linguistic features of those responses were found to be indicative of students' vocabulary knowledge and comprehension skills, which suggests the feasibility of stealth reading assessment using games that embed open-ended questions [26,27]. This study took a different approach by focusing on games implementing multiple-choice questions. Notably, students are less likely to consider the tasks to be multiple-choice questions because they are presented in the context of games. Our results indicate such games can also provide a means to stealthily assess reading skills, which complements the use of NLP methods for reading assessment. In the context of serious games and ITSs, stealth assessment affords ways to evaluate students' literacy skills and update student models as they naturally interact with the software. The stealth assessment of students' reading skills can augment the macro-level adaptation of ITS, such as guiding students' practice across modules. For example, the system may recommend students with lower reading skills to play more summarization games, but direct students with higher reading skills to engage in more difficult practice within the self-explanation module.

Another focus of this study was game enjoyment. Although participants tended to enjoy the games, game enjoyment was associated with reading skills differently in the three games. Reading skill and enjoyment of Vocab Flash were not correlated. However, reading skill was negatively correlated with enjoyment of Dungeon Escape and Adventurer's Loot: participants who enjoyed these two main idea games more had lower scores on the reading skills tests. This result supports the notion that those who are more likely to perform poorly and potentially be frustrated or anxious during a traditional test are also more likely to appreciate playing a game rather than taking a test. On the flip side, participants with higher reading skills may have had more positive experiences taking and succeeding on traditional tests, and thus had less appreciation for the games.

5. Conclusions and Implications

Literacy assessment is a key component of any effort to improve learners' literacy or remediate potential gaps in education. However, such assessments can be slow, disconnected from learning experiences, anxiety-inducing, or boring. The results from this study imply an important practical application as it provides a means to measure learners' literacy skills in real time via games. Stealth assessment via serious games can also inform adaptive instructional paths for students. Serious games and intelligent tutors may act as scaffolding for less skilled readers to receive more personalized instructions to enhance their skills. Furthermore, game-based assessment can replace traditional paper-based literacy measures, which are decontextualized, cause stress and anxiety for test takers, and in turn negatively impact the reliability of these assessments [66–68].

Notably, the games used for stealth assessment in this study were relatively simplistic. Thus, our findings indicate that simple, relatively inexpensive games can be leveraged to assess skills. Nonetheless, more elaborate, immersive games with comparable embedded pedagogical features have strong potential to augment the power of stealth assessment. Note that most participants in this study had Bachelor's or advanced degrees. Future research will involve a broader range of participants with respect to prior education, which will enable assessment of the generalizability of current findings.

Author Contributions: Conceptualization: D.S.M. and Y.F.; Methodology: D.S.M., R.D.R. and Y.F.; Formal analysis: Y.F.; Investigation: Y.F.; Resources: D.S.M.; Data curation: Y.F.; Writing—original draft: Y.F., T.L., L.H. and K.C.; Writing—review and editing: Y.F., T.L., L.H., K.C., R.D.R. and D.S.M.; Supervision: D.S.M. and R.D.R.; Funding acquisition: D.S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Office of Naval Research Grant N00014-20-1-2623 and Institute of Education Sciences Grant R305A190050.

Institutional Review Board Statement: This study was reviewed and approved by ASU's Institutional Review Board, and that the study conforms to recognized standards.

Data Availability Statement: Data may be accessed by emailing the first author at ying.fang07@gmail.com.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results. The opinions expressed are those of the authors and do not represent views of the Office of Naval Research or Institute of Education Sciences.

References

1. Organization for Economic Cooperation and Development. *OECD Skills Outlook: First Results from the Survey of Adult Skills*; OECD Publishing: Paris, France, 2013.
2. NAEP Report Card: Reading. The Nations' Report Card. Available online: https://www.nationsreportcard.gov/reading/nation/achievement?grade=8 (accessed on 20 May 2022).
3. NCES. *Highlights of the 2017 U.S. PIAAC Results Web Report*; Department of Education, Institute of Education Sciences, National Center for Education Statistics: Washington, DC, USA, 2020. Available online: https://nces.ed.gov/surveys/piaac/current_results.asp (accessed on 20 May 2022).
4. Shute, V.J.; Ventura, M. *Measuring and Supporting Learning in Games: Stealth Assessment*; The MIT Press: Cambridge, MA, USA, 2013.
5. Kato, P.M.; de Klerk, S. Serious games for assessment: Welcome to the jungle. *J. Appl. Test. Technol.* **2017**, *18*, 1–6.
6. Francis, D.J.; Snow, C.E.; August, D.; Carlson, C.D.; Miller, J.; Iglesias, A. Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Sci. Stud. Read.* **2006**, *10*, 301–322. [CrossRef]
7. Petroviča, S.; Anohina-Naumeca, A. The adaptation approach for affective game-based assessment. *Appl. Comput. Syst.* **2017**, *22*, 13–20. [CrossRef]
8. Onwuegbuzie, A.J.; Leech, N.L. Sampling Designs in Qualitative Research: Making the Sampling Process More Public. *Qual. Rep.* **2007**, *12*, 238–254. [CrossRef]
9. Kim, Y.J.; Ifenthaler, D. Game-based assessment: The past ten years and moving forward. In *Game-Based Assessment Revisited*; Ifenthaler, D., Kim, Y.J., Eds.; Springer: Cham, Switzerland, 2019; pp. 3–11.
10. O'Connor, M.C.; Paunonen, S.V. Big Five personality predictors of post-secondary academic performance. *Personal. Individ. Differ.* **2007**, *43*, 971–990. [CrossRef]
11. Poropat, A.E. A meta-analysis of the five-factor model of personality and academic performance. *Psychol. Bull.* **2009**, *135*, 322. [CrossRef]
12. Ke, F.; Parajuli, B.; Smith, D. Assessing Game-Based Mathematics Learning in Action. In *Game-Based Assessment Revisited*; Springer: Cham, Switzerland, 2019; pp. 213–227.
13. Min, W.; Frankosky, M.H.; Mott, B.W.; Rowe, J.P.; Smith, A.; Wiebe, E.; Boyer, K.E.; Lester, J.C. DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Trans. Learn. Technol.* **2019**, *13*, 312–325. [CrossRef]
14. Shute, V.J.; Rahimi, S. Stealth assessment of creativity in a physics video game. *Comput. Hum. Behav.* **2021**, *116*, 106647. [CrossRef]

15. Shute, V.; Rahimi, S.; Smith, G.; Ke, F.; Almond, R.; Dai, C.P.; Kuba, R.; Liu, Z.; Yang, X.; Sun, C. Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *J. Comput. Assist. Learn.* **2021**, *37*, 127–141. [CrossRef]
16. Simonson, M.; Smaldino, S.; Albright, M.; Zvacek, S. Assessment for distance education. In *Teaching and Learning at a Distance: Foundations of Distance Education*; Prentice-Hall: Upper Saddle River, NJ, USA, 2000.
17. Shute, V.J.; Leighton, J.P.; Jang, E.E.; Chu, M.W. Advances in the science of assessment. *Educ. Assess.* **2016**, *21*, 34–59. [CrossRef]
18. De-Juan-Ripoll, C.; Soler-Domínguez, J.L.; Guixeres, J.; Contero, M.; Álvarez Gutiérrez, N.; Alcañiz, M. Virtual reality as a new approach for risk taking assessment. *Front. Psychol.* **2018**, *9*, 2532. [CrossRef]
19. De Rosier, M.E.; Thomas, J.M. Establishing the criterion validity of Zoo U's game-based social emotional skills assessment for school-based outcomes. *J. Appl. Dev. Psychol.* **2018**, *55*, 52–61. [CrossRef]
20. Salen, K.; Zimmerman, E. *Rules of Play: Game Design Fundamentals*; The MIT Press: Cambridge, UK, 2004.
21. Tsikinas, S.; Xinogalos, S. Towards a serious games design framework for people with intellectual disability or autism spectrum disorder. *Educ. Inf. Technol.* **2020**, *25*, 3405–3423. [CrossRef]
22. Annetta, L.A. The "I's" have it: A framework for serious educational game design. *Rev. Gen. Psychol.* **2010**, *14*, 105–112. [CrossRef]
23. Wang, L.; Shute, V.; Moore, G.R. Lessons learned and best practices of stealth assessment. *Int. J. Gaming Comput. Mediat. Simul.* **2015**, *7*, 66–87. [CrossRef]
24. Akram, B.; Min, W.; Wiebe, E.; Mott, B.; Boyer, K.E.; Lester, J. Improving stealth assessment in game-based learning with LSTM-based analytics. In Proceedings of the 11th International Conference on Educational Data Mining, Buffalo, NY, USA, 15–18 July 2018.
25. DiCerbo, K.E.; Bertling, M.; Stephenson, S.; Jia, Y.; Mislevy, R.J.; Bauer, M.; Jackson, G.T. An application of exploratory data analysis in the development of game-based assessments. In *Serious Games Analytics*; Springer: Cham, Switzerland, 2015; pp. 319–342.
26. Allen, L.K.; McNamara, D.S. You Are Your Words: Modeling Students' Vocabulary Knowledge with Natural Language Processing. In Proceedings of the 8th International Conference on Educational Data Mining, Madrid, Spain, 26–29 June 2015; Santos, O.C., Boticario, J.G., Romero, C., Pechenizkiy, M., Merceron, A., Mitros, P., Luna, J.M., Mihaescu, C., Moreno, P., Hershkovitz, A., et al., Eds.; International Educational Data Mining Society, 2015; pp. 258–265.
27. Fang, Y.; Allen, L.K.; Roscoe, R.D.; McNamara, D.S. Stealth literacy assessment: Leveraging games and NLP in iSTART. In *Advancing Natural Language Processing in Educational Assessment*; Yaneva, V., Davier, M., Eds.; Routledge: New York, NY, USA, 2023; pp. 183–199.
28. McCarthy, K.S.; Laura, K.A.; Scott, R.H. Predicting Reading Comprehension from Constructed Responses: Explanatory Retrievals as Stealth Assessment. In Proceedings of the International Conference on Artificial Intelligence in Education, Ifrane, Morocco, 6–10 July 2020; Springer: Cham, Switzerland, 2020; pp. 197–202.
29. Li, H. Deep learning for natural language processing: Advantages and challenges. *Natl. Sci. Rev.* **2018**, *5*, 24–26. [CrossRef]
30. Fang, Y.; Li, T.; Roscoe, R.D.; McNamara, D.S. Predicting literacy skills via stealth assessment in a simple vocabulary game. In Proceedings of the 23rd Human-Computer Interaction International Conference, Virtual Conference, 24–29 July 2021; Sottilare, R.A., Schwarz, J., Eds.; Springer: Cham, Switzerland, 2021; pp. 32–44.
31. Freebody, P.; Anderson, R.C. Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Read. Res. Q.* **1983**, *18*, 277–294. [CrossRef]
32. Bernhardt, E. Progress and procrastination in second language reading. *Annu. Rev. Appl. Linguist.* **2005**, *25*, 133–150. [CrossRef]
33. Cain, K.; Oakhill, J. Reading comprehension and vocabulary: Is vocabulary more important for some aspects of comprehension? *L'Année Psychol.* **2014**, *114*, 647–662. [CrossRef]
34. Cromley, J.G.; Azevedo, R. Testing and refining the direct and inferential mediation model of reading comprehension. *J. Educ. Psychol.* **2007**, *99*, 311–325. [CrossRef]
35. Chen, X.; Meurers, D. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *J. Res. Read.* **2018**, *41*, 486–510. [CrossRef]
36. Masrai, A. Vocabulary and reading comprehension revisited: Evidence for high-, mid-, and low-frequency vocabulary knowledge. *Sage Open* **2019**, *9*, 2158244019845182. [CrossRef]
37. Jun Zhang, L.; Bin Anual, S. The role of vocabulary in reading comprehension: The case of secondary school students learning English in Singapore. *RELC J.* **2008**, *39*, 51–76. [CrossRef]
38. Kintsch, W.; Walter Kintsch, C. *Comprehension: A Paradigm for Cognition*; Cambridge University Press: Cambridge, UK, 1998.
39. Brown, A.L.; Campione, J.C.; Day, J.D. Learning to learn: On training students to learn from texts. *Educ. Res.* **1981**, *10*, 14–21. [CrossRef]
40. Bransford, J.D.; Brown, A.L.; Cocking, R.R. *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*; National Academy Press: Washington, DC, USA, 2000.
41. Wade-Stein, D.; Kintsch, E. Summary Street: Interactive computer support for writing. *Cogn. Instr.* **2004**, *22*, 333–362. [CrossRef]
42. Fox, E. The Role of Reader Characteristics in Processing and Learning from Informational Text. *Rev. Educ. Res.* **2009**, *79*, 197–261. [CrossRef]
43. Williams, J.P. Teaching text structure to improve reading comprehension. In *Handbook of Learning Disabilities*; Swanson, H.L., Harris, K.R., Graham, S., Eds.; The Guilford Press: New York, NY, USA, 2003; pp. 293–305.

44. Anmarkrud, Ø.; Bråten, I.; Strømsø, H.I. Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learn. Individ. Differ.* **2014**, *30*, 64–76. [CrossRef]
45. Wigent, C.A. High school readers: A profile of above average readers and readers with learning disabilities reading expository text. *Learn. Individ. Differ.* **2013**, *25*, 134–140. [CrossRef]
46. Lau, K.L. Reading strategy use between Chinese good and poor readers: A think aloud study. *J. Res. Read.* **2006**, *29*, 383–399. [CrossRef]
47. Shores, J.H. Are fast readers the best readers? A second report. *Elem. Engl.* **1961**, *38*, 236–245.
48. Johnston, P.; Afflerbach, P. The process of constructing main ideas from text. *Cogn. Instr.* **1985**, *2*, 207–232. [CrossRef]
49. Afflerbach, P.P. The influence of prior knowledge on expert readers' main idea construction strategies. *Read. Res. Q.* **1990**, *25*, 31–46. [CrossRef]
50. Chittaro, L.; Buttussi, F. Exploring the use of arcade game elements for attitude change: Two studies in the aviation safety domain. *Int. J. Hum. Comput. Stud.* **2019**, *127*, 112–123. [CrossRef]
51. Derbali, L.; Frasson, C. Players' motivation and EEG waves patterns in a serious game environment. In *International Conference on Intelligent Tutoring Systems*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 297–299.
52. Jackson, G.T.; McNamara, D.S. Motivation and performance in a game-based intelligent tutoring system. *J. Educ. Psychol.* **2013**, *105*, 1036. [CrossRef]
53. McNamara, D.S. Self-explanation and reading strategy training (SERT) improves low-knowledge students' science course performance. *Discourse Process.* **2017**, *54*, 479–492. [CrossRef]
54. McCarthy, K.S.; Watanabe, M.; Dai, J.; McNamara, D.S. Personalized learning in iSTART: Past modifications and future design. *J. Res. Technol. Educ.* **2020**, *52*, 301–321. [CrossRef]
55. McNamara, D.S.; Boonthum, C.; Levinstein, I.B.; Millis, K. Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In *Handbook of Latent Semantic Analysis*; Landauer, T., McNamara, D.S., Dennis, S., Kintsch, W., Eds.; Erlbaum: Hillsdale, MI, USA, 2007; pp. 227–241.
56. McNamara, D.S. Chasing theory with technology: A quest to understand understanding. *Discourse Process.* **2021**, *58*, 442–448. [CrossRef]
57. VanLehn, K. The behavior of tutoring systems. *Int. J. Artif. Intell. Educ.* **2006**, *16*, 227–265.
58. Shute, V.J.; Psotka, J. Intelligent Tutoring Systems: Past, Present and Future. In *Handbook of Research on Educational Communications and Technology*; Jonassen, D., Ed.; Macmillan: New York, NY, USA, 1996; pp. 570–600.
59. Woolf, B.P. *Building Intelligent Interactive Tutors: Student-Centered Strategies for Revolutionizing E-Learning*; Morgan Kaufmann: San Francisco, CA, USA, 2010.
60. Phillips, L.M.; Norris, S.P.; Osmond, W.C.; Maynard, A.M. Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. *J. Educ. Psychol.* **2002**, *94*, 3–13. [CrossRef]
61. Davies, M. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *Int. J. Corpus Linguist.* **2009**, *14*, 159–190. [CrossRef]
62. Kimura, T. The impacts of computer adaptive testing from a variety of perspectives. *J. Educ. Eval. Health Prof.* **2017**, *14*, 1149050. [CrossRef]
63. Hulin, C.; Netemeyer, R.; Cudeck, R. Can a reliability coefficient be too high? *J. Consum. Psychol.* **2001**, *10*, 55–58.
64. McClarty, K.L.; Orr, A.; Frey, P.M.; Dolan, R.P.; Vassileva, V.; McVay, A. *A Literature Review of Gaming in Education*; Research Report; Pearson: Hoboken, NJ, USA, 2012; Available online: https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/tmrs/lit-review-of-gaming-in-education.pdf (accessed on 22 May 2023).
65. Shute, V.J.; Rahimi, S. Review of computer-based assessment for learning in elementary and secondary education. *J. Comput. Assist. Learn.* **2017**, *33*, 1–19. [CrossRef]
66. Cassady, J.C.; Johnson, R.E. Cognitive test anxiety and academic performance. *Contemp. Educ. Psychol.* **2002**, *27*, 270–295. [CrossRef]
67. Segool, N.K.; Carlson, J.S.; Goforth, A.N.; Von Der Embse, N.; Barterian, J.A. Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychol. Sch.* **2013**, *50*, 489–499. [CrossRef]
68. Von der Embse, N.P.; Witmer, S.E. High-stakes accountability: Student anxiety and large-scale testing. *J. Appl. Sch. Psychol.* **2014**, *30*, 132–156. [CrossRef]