# Stealth Assessment of Reading Comprehension Skill and Vocabulary Knowledge using Game Performance

by

Katerina Christhilf

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved August 2023 by the
Graduate Supervisory Committee:

Danielle McNamara, Chair
Rod Roscoe
Gene Brewer

ARIZONA STATE UNIVERSITY

December 2023

# ABSTRACT

The current study explores the extent to which literacy game performance can be used to assess reading comprehension skill and vocabulary knowledge. Standardized reading assessments have the benefit of years of validation across different age groups and reading comprehension levels, allowing teachers to evaluate students' reading performance and relate it to a national standard. However, these assessments reduce classroom time for learning activities, which may be more authentic indicators of student progress. Students' reading skills can be measured during learning activities by using game-based stealth assessment of literacy. Game-based assessment may be more enjoyable and less likely to invoke test anxiety than traditional assessments, but enjoyment may also impact the validity of the assessment. The current study recruited participants (n=405) to play five literacy games: CON-Artist, Paraphrase Quest, Fix It, Map Conquest, and Vocab Flash. Students also completed the Gates-MacGinitie Reading Test (GMRT), which serves as a validated measure of reading comprehension skill and vocabulary knowledge. Students answered enjoyment questions after each game and the GMRT, and they completed the Cognitive Test Anxiety questionnaire, which measures trait-level negative thoughts about test-taking. The results indicate that Vocab Flash predicted 31% of variance in reading comprehension and 21% of variance in vocabulary knowledge. The other games were not predictive beyond Vocab Flash, but each of them was weakly correlated with reading comprehension skill and vocabulary knowledge. Three games were more enjoyable than GMRT Reading Comprehension, but no games were more enjoyable than GMRT Vocabulary. Cognitive Test Anxiety was negatively correlated with the GMRT and Vocab Flash, but not with the other games. Game enjoyment moderated the relationship between game performance and reading skill, albeit in differing directions. Paraphrase Quest was less predictive of reading comprehension for students who enjoyed the game, and Vocab Flash was more predictive of reading comprehension for those who enjoyed the game. The findings of this study suggest that a simple vocabulary game can be used to measure reading comprehension skill and vocabulary knowledge. Future research is needed to better understand how game-based assessments can be designed to minimize the effects of test anxiety and enjoyment on performance.

# TABLE OF CONTENTS

Page

APPENDIX Page

# LIST OF TABLES

# LIST OF FIGURES

Figure Page

# CHAPTER 1

# INTRODUCTION

Reading is a ubiquitous task, yet it is also a complicated skill that takes years to acquire. Much of student learning occurs from textbooks, making reading comprehension skills essential for academic success. University entrance exams include challenging reading components, and employers seek workers with strong language skills, including reading. However, few people in the U.S. are highly skilled readers, and many others have little to no reading skills. Approximately 43 million U. S. adults have low or no English skills (NCES, 2019). About 30% of U.S. 12th grade students do not have basic reading proficiency (NAEP, 2019), indicating that they do not adequately comprehend texts. Hence, more research is needed to understand how to help students achieve reading proficiency, so that all students are prepared for the demands of modern life. Precise reading assessments are necessary to estimate students' current skills and provide instruction to improve assessed deficiencies.

Many traditional reading assessments are not optimal for promoting student learning. Reading assessments may involve reading short passages and answering questions, providing missing words, or re-telling a story (Schrank & Wendling, 2018; Smolkowski & Cummings, 2015). These tasks may be inauthentic, because they do not match the real-life situations in which reading occurs or knowledge obtained from reading is used. The graded questions occur after reading, so they do not measure moment-to-moment processes that might explain when and how a reader is failing to comprehend the text (Magliano et al., 2007). Additionally, reading assessments may reduce classroom time for learning activities and limit how often students can be tested.

For example, the standard Gates-MacGinitie Reading Test (GMRT) used in this study requires 55 minutes to administer (*Gates-MacGinitie Reading Tests*, 1989). Infrequent testing in turn renders it difficult to give students individualized feedback. Traditional assessments may also not clarify the specific competencies on which students need to improve (Snyder et al., 2005). Furthermore, some students have testing anxiety, so their scores may not accurately reflect their current reading skills (Cassady & Johnson, 2002). Stealth assessment is a potential solution to these problems.

Stealth assessment refers to testing students during a learning activity, such that students are unaware of being tested (Shute, 2011). Traditionally, learning activities are often separate from testing: learning is provided in a lecture format, followed by graded tests of learning. Stealth assessment combines learning and assessment, resulting in less time spent testing and more immediate feedback to students. Digital learning activities can be designed to assess students' current skills through multiple choice problems, open-response activities, or even mouse movements. Learning activities can provide automated feedback, adapt to the learner's skill level, and recommend different activities based upon the student's weaknesses by tracking the student's current skills. Electronic stealth assessment can give teachers quick estimates of students' progress that are directly tied to the reading comprehension competencies present in student learning activities. Stealth assessment also may be more enjoyable and engaging, which increases the validity of the assessment (Klimmt et al., 2009).

The current study focuses on two aspects of literacy: reading comprehension skill and vocabulary knowledge. The goal of the study is to examine the extent to which stealth assessment in games can predict reading comprehension skill and vocabulary knowledge, as estimated by standardized tests. The study also aims to explore the extent to which game enjoyment, test anxiety, and technology acceptance affect the predictive validity of game performance. The games in the current study are chosen from iSTART, the interactive Strategy Training for Active Reading and Thinking. iSTART is an interactive trainer that teaches students to use reading strategies through videos, animated agents, games, and self-explanation practice. It has been demonstrated to improve reading comprehension and self-explanation skills (McNamara, 2017). iSTART games provide practice in strategy identification, self-explanation, and synonym identification. They also have a strong potential to serve as stealth literacy assessments by measuring aspects of reading comprehension skill and vocabulary knowledge. Exploring which types of games are more accurate measures of literacy skills could inform future game development as well as classroom assessments. The current study uses five games from iSTART: Vocab Flash, CON-Artist, Paraphrase Quest, Fix It, and Map Conquest.

The following sections provide an overview of the process of reading comprehension and the tasks that can be used to improve and measure it. These descriptions are followed by a discussion of the development and use of game-based stealth assessment.

# CHAPTER 2

# LITERATURE REVIEW

## Reading Comprehension

Reading comprehension is the process of constructing meaning from text. The Construction-Integration Model (Kintsch, 1988, 1998) suggests that readers form a mental representation of the text based on explicit text material, inferences, and prior knowledge. Comprehension constructed solely from explicit text material, including forming connections across the text, is referred to as the *textbase*. Integrating the textbase with one's prior knowledge, creating a more elaborated mental model, is referred to as forming the *situation model*. A successful reader goes through cycles of constructing the textbase and integrating prior knowledge to form the situation model. Each sentence that is read is informed by the context of prior sentences. The textbase and the situation model are formed at two levels of text: the microstructure and the macrostructure. The microstructure refers to the sentence level of the text, while the macrostructure refers to the global organization of the text. The reader's mental representation of the macrostructure may emerge from the hierarchical organization of information from the microstructure or from more explicit signifiers in the text, such as headers (McNamara & Magliano, 2009).

Reading comprehension is therefore a complex process, ranging from decoding individual words to comprehending sentences to understanding the overall structure and meaning of the text. Reading comprehension skill can be measured through standardized comprehension questions, but there may be a benefit to using a variety of tasks to assess comprehension. Different tasks might target more specific knowledge and tasks that relate to successful reading comprehension. They can also serve as both learning activities and assessment at the same time. These include selection of synonyms and connective words to measure vocabulary knowledge, paraphrasing, self-explanation, and text summarization.

**Vocabulary.** The current study uses the game Vocab Flash to assess vocabulary knowledge; students are asked to select the best synonym for the target word. Students build vocabulary primarily through reading (Anderson & Nagy, 1992), which means that vocabulary knowledge and reading comprehension are inextricably linked. Students with greater depth of vocabulary knowledge tend to be better at reading comprehension (Ouelette, 2006). Vocabulary knowledge enhances readers' ability to produce thematically related inferences across a text to comprehend the text's underlying meaning (Cain & Oakhill, 2014). Readers are better able to build connections across the text when they have a deep understanding of the individual words and the concepts associated with them. Understanding these words allows readers to recognize themes in the text that are not explicitly stated. Thus, depth of vocabulary knowledge is an important factor in being able to answer global cohesion questions about a text.

**Connectives.** The current study uses the game CON-Artist to assess knowledge of connectives; students are asked to select the best word or phrase to connect two given sentences. Connectives can be considered a special subset of vocabulary knowledge that is particularly important to text comprehension. Connectives give readers clues about how different sentences or clauses relate to each other, helping them integrate successive clauses. Connecting ideas from different parts of the text helps readers form a coherent mental representation of a text's macrostructure. Connectives might signal an additional supportive idea, a temporal relation between two events, one event causing another event, or an idea opposing the previously given idea. For example, the connective “however” allows one to infer that the sentence will provide an opposing argument to the preceding sentence. Cognitive load and processing times are reduced when sentences contain effective connectives (Millis & Just, 1994; van Silfhout et al., 2015). Knowledge of connectives accounts for variance in text comprehension performance above and beyond general breadth of vocabulary (Crosson & Lesaux, 2011; Kohnen & Retelsdorf, 2019) and receptive grammar (Volodina et al., 2021).

**Paraphrasing.** The current study uses the game Paraphrase Quest to assess paraphrasing skill; students are asked to select the best paraphrase of a given sentence. Rephrasing sentences in one’s own words, or paraphrasing, is an essential and common process during reading. While paraphrasing is considered a surface-level strategy, it is nonetheless an important first step to comprehending the text. When reading a sentence, a successful reader converts the explicit words in the sentence to a mental representation of the sentence’s meaning. That process often requires paraphrasing, so that one can more readily comprehend the meaning of the sentence. Young, developing readers who are trained to paraphrase tend to have improved reading comprehension skills (Hagaman et al., 2016). Likewise, students who are better at paraphrasing are more likely to comprehend the text (Haynes & Fillmer, 1984).

**Self-explanation.** The current study asks students to produce self-explanations in the game Map Conquest; the quality of the self-explanations is scored by iSTART. One way to encourage the use of more deliberate comprehension of the text is to ask students to explain the text in their own words, referred to as self-explanation. Self-explanation prompts the reader to actively generate ideas and connections between ideas (McNamara, 2009). Readers are encouraged to explain the text using both text and their own knowledge, instead of passively reading.

Students' self-explanations in Map Conquest are processed using natural language processing to examine their length, cohesion, and lexical diversity. Self-explanations can be analyzed through human coding or through natural language processing. The quality of self-explanations can be used to estimate a student's comprehension of the text (Allen, Snow, & McNamara, 2015; Boonthum-Denecke et al., 2011; Ozuru et al., 2010; Sinclair et al., 2021). Skilled readers tend to use more sophisticated words and more complex syntax (Crossley et al., 2015; McNamara et al., 2010). They also typically have higher semantic overlap between successive self-explanations (Graesser & McNamara, 2011).

**Summarizing.** In the game Fix It, students are asked to determine if a summary is good, and if not, to identify the best way to fix it. A summary refers to a condensed version of a text that includes the main ideas without minor details. Writing an effective summary means condensing a text into its most important ideas. It provides an opportunity to reflect on the text, examine links between ideas, and distill the overall message or theme. Students who are trained to summarize texts tend to do better on standardized measures of reading comprehension. Based on a recent meta-analysis of 19 studies (Graham & Herbert, 2010), the average weighted effect size of writing summaries on text comprehension is 0.52, compared to controls such as re-reading text, studying text, and receiving reading instruction. Summarization training is thought to help by requiring students to connect different parts of the text to find the "gist". Summarization training also helps students monitor their comprehension, as they prepare to consolidate their understanding of the text in a summary. The process of re-working a summary forces a reader to think deeper about the text to try to find its underlying meaning.

The quality of a summary can indicate how well a student has understood the text. Summaries are best written when one first builds an accurate mental model of the text, then condenses it into the most important ideas. Good readers are more likely to remember essential parts of the text, which are ideas that are tied to other ideas in the text (van den Broek et al., 2012). Less successful readers do not discriminate as well between details and main ideas. Therefore, being able to distinguish between good and poor summaries might provide insight into a student's text comprehension.

## Game-Based Stealth Assessment

Stealth assessment refers to measuring skills during a learning activity, such that the learner is not aware of being tested. The goal when designing stealth assessment is to avoid giving learners a feeling of being tested, yet still successfully measure specific skills, so that the stealth assessment could serve as an estimate of performance on a standardized assessment. Many students struggle with test anxiety, and in some cases test anxiety might negatively impact a student's test performance (von der Embse et al., 2018). Measuring skills without the learner's awareness reduces test anxiety, which might improve the reliability and validity of the assessment.

Assessments can be made more authentic by assessing literacy skills with a learning activity instead of an isolated test, such as by simulating real-world experiences. The learning process is not interrupted for assessment, and classroom time spent assessing students is reduced. Using stealth assessment in an online activity can further improve the learning process when paired with automated, adaptive feedback. Literacy skills can be quickly measured, followed by immediate, personalized feedback and adaptive changes to the learning activity. Such feedback can be difficult to provide in large classrooms, especially with standardized tests that are not subdivided into learning competencies (Snyder et al., 2005).

Using games as stealth assessment can make the learning process less stressful, more enjoyable, and more engaging. High engagement is essential for students to adequately gain new knowledge and skills from practice activities (Arum & Roksa, 2011; Parsons & Taylor, 2011). Numerous meta-analyses have suggested that participants find games to be more engaging than traditional learning activities and standardized tests (Barab et al., 2012; Wilson et al., 2009). A well-designed, engaging game may make the assessment unnoticeable to participants, while similar non-game learning activities would make the assessment more obvious. One can also measure more distinct aspects of reading comprehension by using a variety of games, providing reading sub-scores that are more informative to students than a traditional test.

Stealth assessment has successfully been used to measure a variety of learning competencies. A game called Use Your Brainz, modeled on Plants vs. Zombies II, measures reasoning ability based on how students place different plans (Shute et al., 2016). The game Newton's Playground measures conceptual physics understanding, persistence, and creativity (Shute et al., 2013). Externally validated measures of these competencies are correlated with Newton's Playground performance, and game practice is associated with improvement on an external physics test. In terms of literacy skills, stealth assessment has previously been used within a vocabulary game known as Vocab Flash. Game performance in Vocab Flash accounted for 76% of variance in a vocabulary test, and 74% of variance in a reading comprehension test (Fang et al., 2021). Stealth assessment can also be incorporated into generative games, which are games in which participants produce written responses. The quality of students' written responses can be analyzed using natural language processing to reflect a students' understanding of a text (Allen, Snow, & McNamara, 2015).

One approach that has been used to design stealth assessment is known as evidence-centered design (Mislevy et al., 2003). According to this perspective, designers should first determine the specific learning competencies they wish to measure, such as vocabulary knowledge and comprehension skill. In turn, it is necessary to delineate what constitutes valid evidence of the competency, such as being able to identify synonyms of given words. Finally, a game is designed that includes tasks that elicit such evidence. For example, when students play Vocab Flash, their performance on the evidence-giving task is recorded into a database. The game can be programmed to give students feedback in response to their performance, such as displaying whether their response was correct, providing examples, and so on. The game can also adapt according to the participant's responses by making the game more challenging when a student is performing well or making the game easier when a student is performing poorly. For instance, Vocab Flash has words leveled by difficulty. Users are given more challenging words after repeated correct responses, and less challenging words after repeated incorrect responses. Incorporating such adaptivity into a game allows the game to more precisely assess the student's vocabulary level while providing an optimum level of difficulty to the student.

# CHAPTER 3

# CURRENT STUDY

The current study aims to examine the extent to which iSTART game performance provides measures of comprehension skill and vocabulary knowledge, as measured by multiple choice assessments. Five literacy games have been selected for participants to play. Four of these games are identification games, which are games that ask students to select a choice out of several options. One game, Map Conquest, is a generative game that asks students to produce a written response. These games measure various aspects of literacy knowledge and skills: vocabulary knowledge, summarization skill, knowledge of connectives, self-explanations, and paraphrasing. Reading comprehension relies on a variety of competencies, so using a variety of games could account for these different dimensions, perhaps better than a traditional reading test. A single game might be sufficient to predict test performance if the games measure similar aspects of reading comprehension. If each game accounts for unique variance, this would support the benefits of using a variety of games to capture reading comprehension and vocabulary knowledge. This study fills a gap in the research by determining whether using a variety of games, both identification and generative, provides more accurate stealth assessment than a single game.

## Primary Research Questions and Hypotheses

**RQ1.** Does student performance in literacy games accurately predict performance in standardized literacy assessments? The study aims to examine the extent to which five literacy games collectively predict performance on the GMRT.

**H1.** Student performance in literacy games will be an effective estimate of performance in standardized literacy assessments. Map Conquest, Fix It, Vocab Flash, Paraphrase Quest, and CON-Artist performance will account for variance in GMRT Reading Comprehension and Vocabulary. Prior work has demonstrated that games can successfully be used as proxies for standardized test performance (Fang et al., 2021; Shute et al., 2016).

**RQ1a.** Does each game account for unique variance in GMRT Reading Comprehension performance? Each game relates to a specific component of reading skill. Vocab Flash measures players' vocabulary knowledge by asking them to name synonyms of words. Paraphrase Quest measures paraphrasing ability by asking players to choose the best paraphrase of a sentence. Fix It measures summarization abilities by asking players to determine how a summary is broken. Map Conquest measures self-explanation ability, by asking players to self-explain a text at various points. CON-Artist measures knowledge and use of connectives by asking players to determine which connective fits between two sentences.

**H1a.** Each game will account for some unique variance in GMRT Reading Comprehension, as each game measures a different reading competency, and each competency is partially distinct and related to reading comprehension. Such a result would suggest the benefits of using multiple games to measure overall reading comprehension, rather than a single game. The games could also be used individually to indicate specific aspects of reading comprehension with which a student is succeeding or struggling.

Vocab Flash and CON-Artist will account for the most variance in Reading Comprehension, given that in a prior study Vocab Flash accounted for 74% of variance (Fang et al., 2021). Fix It and Paraphrase Quest will account for a small but significant amount of variance in test performance. Map Conquest natural language processing indices will account for the next most variance, as in a prior study linguistic indices accounted for 27% of variance (Allen, Snow, & McNamara, 2015).

**RQ1b.** Does each game account for unique variance in GMRT Vocabulary performance? The predictive capacities of the games may differ for vocabulary test performance, compared to reading comprehension test performance.

**H1b.** Vocab Flash will account for the most variance in vocabulary, given that in a prior study Vocab Flash accounted for 76% of variance (Fang et al., 2021). The game Vocab Flash is specifically designed to elicit vocabulary knowledge. Performance in Vocab Flash can therefore be expected to account for variance in GMRT Vocabulary performance. However, the GMRT Vocabulary section contextualizes the words in a phrase, so other reading skills such as paraphrasing may be involved in test performance. If other games do account for additional variance in vocabulary performance, this suggests the need for a more nuanced approach to measuring vocabulary knowledge than using a single game such as Vocab Flash. Map Conquest natural language processing indices, particularly lexical sophistication, will account for the next most variance. Fix It and Paraphrase Quest will account for a small but significant amount of variance in test performance.

## Secondary Research Questions and Hypotheses

**RQ2a.** Is a person's enjoyment of literacy games predictive of literacy test performance above and beyond game performance? Game mechanics could make learning and testing more engaging and motivating, as well as more authentic if the games mimic some real-life situations. However, some students might find the games more enjoyable than others, and this enjoyment or dislike could affect performance. If so, the game mechanics may reduce the validity of the stealth assessment. Alternatively, students who perform well on a game might find it more enjoyable due to their performance on the game.

**H2a.** Enjoyment of literacy games will not account for variance in GMRT performance above and beyond literacy game performance. Game accuracy is expected to be the sole predictor of literacy test performance, with enjoyment being related to literacy test performance only to the extent that it is related to game performance. Using these games as stealth assessment would require assessing both accuracy and enjoyment if game enjoyment is an additional significant predictor.

**RQ2b.** Is a self-explanation game, Map Conquest, as predictive of reading test performance as a comparable non-game task, the Self-Explanation Task? Both measures ask students to read a text and self-explain target sentences at key points. The dependent variables are selected natural language processing indices. Map Conquest also includes a game component that is displayed after each self-explanation: placing flags on a map to try to conquer all the territories. Map Conquest will account for the same amount of variance in GMRT performance as the Self-Explanation Task if game mechanics do not affect the predictive validity of the game. Alternatively, game mechanics may render

Map Conquest more engaging, in turn increasing validity, or alternatively, more difficult, in turn distracting some participants from the literacy aspects of the game and decreasing validity.

**H2b.** Map Conquest and the Self-Explanation Task will account for equal variance in GMRT performance. Map Conquest is very similar to the Self-Explanation Task, except for the game components in Map Conquest. Both involve reading a text and self-explaining it at key points. If Map Conquest and the Self-Explanation Task predict the same amount of variance in GMRT performance, then the game features used in Map Conquest do not reduce the predictive validity of the game.

**RQ2c.** Do students enjoy literacy games more than non-game literacy assessments? One potential advantage of using literacy games as stealth assessment is that students perceive games as more enjoyable than traditional assessments. However, many literacy games are similar to literacy assessments, in that they require reading texts and answering questions. While games in general are more enjoyable than assessments, it is unclear whether this increased enjoyment extends to literacy games. Game enjoyment might not necessarily be higher for literacy games than literacy tests.

**H2c.** Students will enjoy each of the literacy games more than each of the literacy assessments. The games include non-literacy components that are designed to be enjoyable, such as mini-sudokus and placing flags to conquer map territories. These components are expected to make the games more enjoyable than literacy assessments.

**RQ2d.** Is test anxiety, as measured by a test anxiety scale, more related to test performance than game performance?

**H2d.** Test anxiety will be negatively correlated with GMRT performance, but test anxiety will not be correlated with game performance. High test anxiety is associated with low test performance (von der Embse et al., 2018). Test anxiety refers to worried thoughts, often accompanied by physiological responses, in reaction to evaluative situations, particularly academic tests. A standardized test, such as the GMRT, has a similar format to classroom tests. The format makes it more obviously an evaluation of literacy skills than the games. Therefore, test anxiety could be expected to negatively impact GMRT performance more than game performance.

# CHAPTER 4

# METHODS

## Participants

Participants were Arizona State University undergraduate psychology students with basic English proficiency. Participants were compensated with 2 Sona credits. A total of 570 students participated in the study. Of these students, 46 were removed due to missing data and 119 were removed due to inattentive participation. Many students (n = 117) completed all key measures except for CON-Artist, perhaps due to the more difficult instructions preceding game play. Therefore, CON-Artist was removed from all analyses. The final sample consists of 405 participants, of which 184 participated in the fall semester and 221 participated in the spring semester. The participants were 59% men, 22% first-generation college students, and 15% learned English as a second language. The average age was 19, and participants were 54% white, 17% Hispanic, 14% Asian, 6% black or African American, 6% multiracial, and 1% American Indian or Alaska native.

## Materials

**Demographics.** Participants are asked questions about their sex, gender, age, ethnicity, education, native language, and use of English. See Appendix A for a list of the questions.

**Socio-emotional assessments.**

***Technology Acceptance Model.*** The Technology Acceptance Model (TAM) Questionnaire gauges participants' comfort level with using technology (Davis, 1989).

The TAM is included to examine whether technology acceptance is correlated with game performance. See Appendix B for the questionnaire.

***Cognitive Test Anxiety Scale.*** Test anxiety involves a reduction in test concentration due to worry and dread about one's performance, which is associated with negative test performance outcomes. The Cognitive Test Anxiety Scale is designed to measure the cognitive aspects of students' test anxiety, such as lack of confidence and worries about failure. The internal reliability is $\alpha = .86$, and high anxiety on the measure is correlated with poor performance in classroom examinations and the Scholastic Aptitude Test (Cassady & Johnson, 2002). See Appendix C for the Cognitive Test Anxiety Scale.

### Reading skills assessments.

***Self-Explanation Task.*** The Self-Explanation Task is used as a comparison to one of the study games, Map Conquest. These two tasks are similar, excepting the game mechanics in Map Conquest. The comparison is made to explore whether a non-game or game format is more conducive to assessing literacy skills. Participants are prompted to self-explain a scientific text, followed by comprehension questions.

The instructions direct students to self-explain a passage as they read, using strategies such as paraphrasing, bridging, and elaborating. They are provided with examples of self- explanations for each strategy type. The passage participants read is called "Red Blood Cells", and it is a scientific text describing the structure and function of red blood cells (words = 281, paragraphs = 4, Flesch Reading Ease = 56.1, Flesch-Kincaid grade = 8.9). This passage was selected to represent the type of scientific text a student might be asked to comprehend in school. Participants are prompted to type a self-explanation at 9 points during the passage; the sentence to be self-explained is bolded. They are only shown the passage up to the bolded sentence. This task is intended to determine whether such an exercise is differentially predictive from games that similarly prompt self-explanations.

After reading the passage, participants are taken to a new survey page, so that they can no longer view the passage. They are asked 8 short-answer questions that test their comprehension of the passage. Four questions require use of explicit information in the text, and four questions require an inference between different parts of the text. This task serves as a measure of reading comprehension. The text is removed from the page, so participants answer by recalling the information they gleaned from the text. The assessment therefore tests both comprehension and retrieval processes. This design is intended to mimic real-life practice of reading, in which information often is comprehended, stored, and retrieved in the absence of the text (such as when taking a test, conversing with colleagues, etc.).

The self-explanations were analyzed using natural language processing tools, described later. Participants' responses to the comprehension questions were coded for accuracy according to an established rubric (Ozuru et al., 2013; Taylor et al., 2006). The text and questions have been used in a prior study (McCarthy et al., 2018), with raters achieving good reliability in scoring the comprehension questions (Cohen's Kappa = .85). See Appendix D for the text, self-explanation prompt, and questions.

***Gates-MacGinitie Reading Test (GMRT).*** The Gates-MacGinitie Reading Test (4th ed.; Form T, level 10/12; *Gates-MacGinitie Reading Tests*, 1989) is used to assess reading comprehension and vocabulary via two distinct sub-scores. The GMRT has been previously validated as a measure of reading skill (α =.85–.92; Phillips et al., 2002). In the comprehension section, participants read passages that are 3 to 14 sentences long. After each passage, they answer two to six multiple choice questions that measure comprehension of shallow and deep level information. Participants have access to the passage as they answer questions, reducing the memory load of the test. The test is timed such that participants have 20 minutes to complete the 48 questions. In the vocabulary section, students are shown underlined vocabulary words, each embedded in a sentence or phrase. They are asked to choose the closest synonym to the word from a list of five. Participants are timed and given 7 minutes to complete the questions. See Appendix E for the comprehension and vocabulary questions.

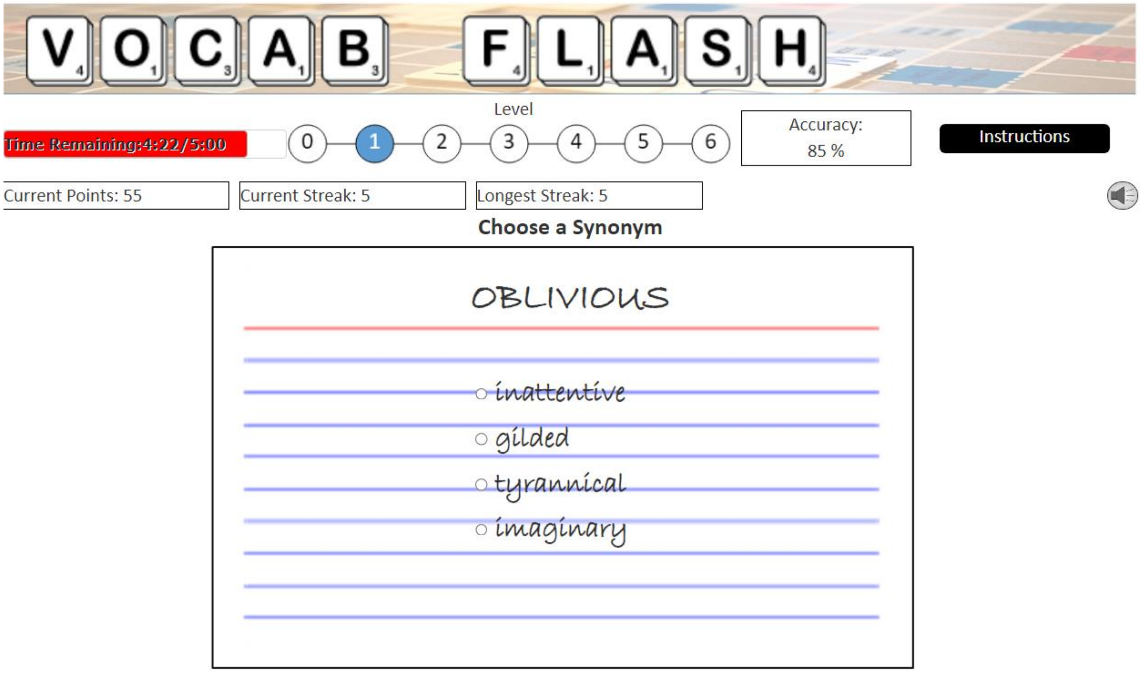
**User experience.**

***Enjoyment and Feedback Survey.*** This survey is given after the GMRT Reading Comprehension section, the GMRT Vocabulary section, the Self-Explanation Task, and each game. Participants are asked a few brief questions about their enjoyment of the measure, their perceived difficulty in completing the measure, and any ways it could be improved. These questions serve to ask participants about their immediate reactions to the measures. The questions are used to examine the extent to which game enjoyment accounts for variance in literacy test performance above and beyond game performance. They are also used to compare enjoyment across tasks. See Appendix F for the full survey.

***User Experience Survey.*** Participants are given this survey at the end of the study to provide their overall impressions about the games. They are asked questions about their enjoyment and attention during the games. They are also asked to provide feedback on the usability of the games and any ways the game could be improved. See Appendix G for the full survey.

**Games.**

***Vocab Flash.*** This game is included to measure vocabulary knowledge. Fang et al. (2021) reported that Vocab Flash accounts for 74% of variance in the GMRT Reading Comprehension section, and 76% of variance in the GMRT Vocabulary section. Users are given five minutes to complete as many flashcards as possible. Each flashcard contains a word and four potential synonyms. Participants choose which word is closest in meaning to the top word. There are seven levels of flashcards. A participant progresses to the next level if they are maintaining a high accuracy in their responses. They move down a level if their accuracy begins to decrease. The dependent variables are the overall proportion of correct responses, the highest level achieved, and average reaction times.



Time Remaining:4:22/5:00

Accuracy: 85 %

Current Points: 55

Current Streak: 5

Longest Streak: 5

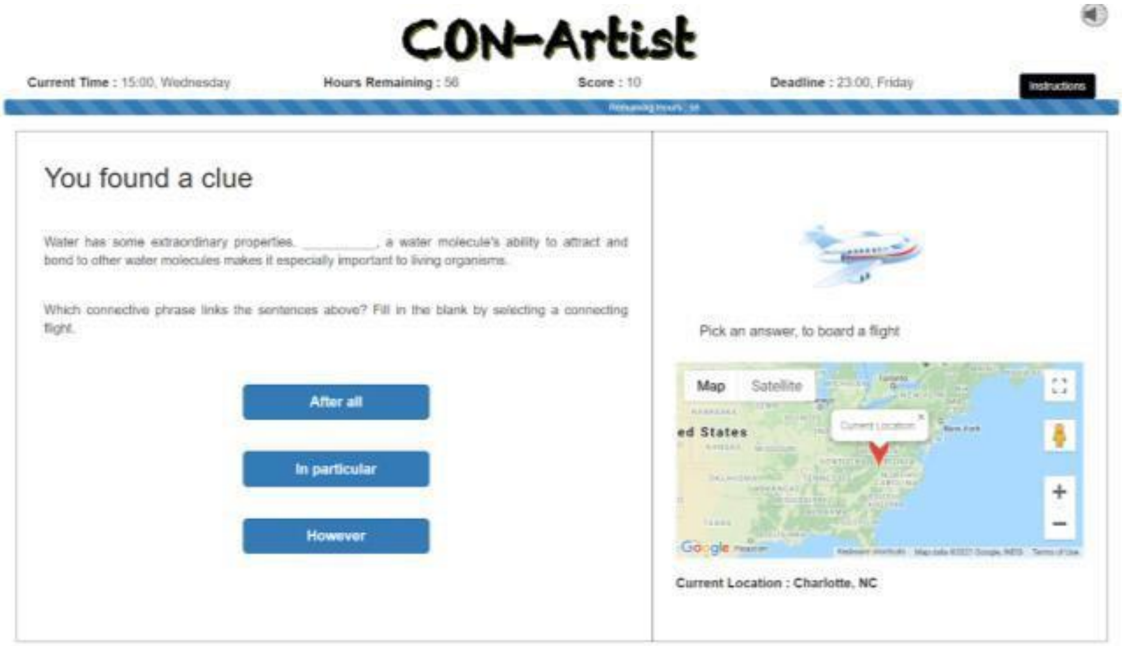
Choose a Synonym

OBLIVIOUS

○ inattentive
○ gilded
○ tyrannical
○ imaginary

***Figure 1.*** Vocab Flash game. This is one of the flashcards that participants might see at Level 1.

***CON-Artist.*** CON-Artist is intended to measure how well participants understand the relations between sentences and the use of connecting words. Participants are asked to choose which transition word best connects two given sentences. The narrative is that a criminal has escaped, and participants are travelling the world to find him. The participant boards the correct plane to the next clue if they respond correctly. If the response is incorrect, the participant boards the wrong plane and is sent back. The dependent variables are the proportion of correct answers and the average reaction time.



CON-Artist

**Current Time :** 15:00, Wednesday **Hours Remaining :** 56 **Score :** 10 **Deadline :** 23:00, Friday

You found a clue

Water has some extraordinary properties. __________, a water molecule's ability to attract and bond to other water molecules makes it especially important to living organisms.

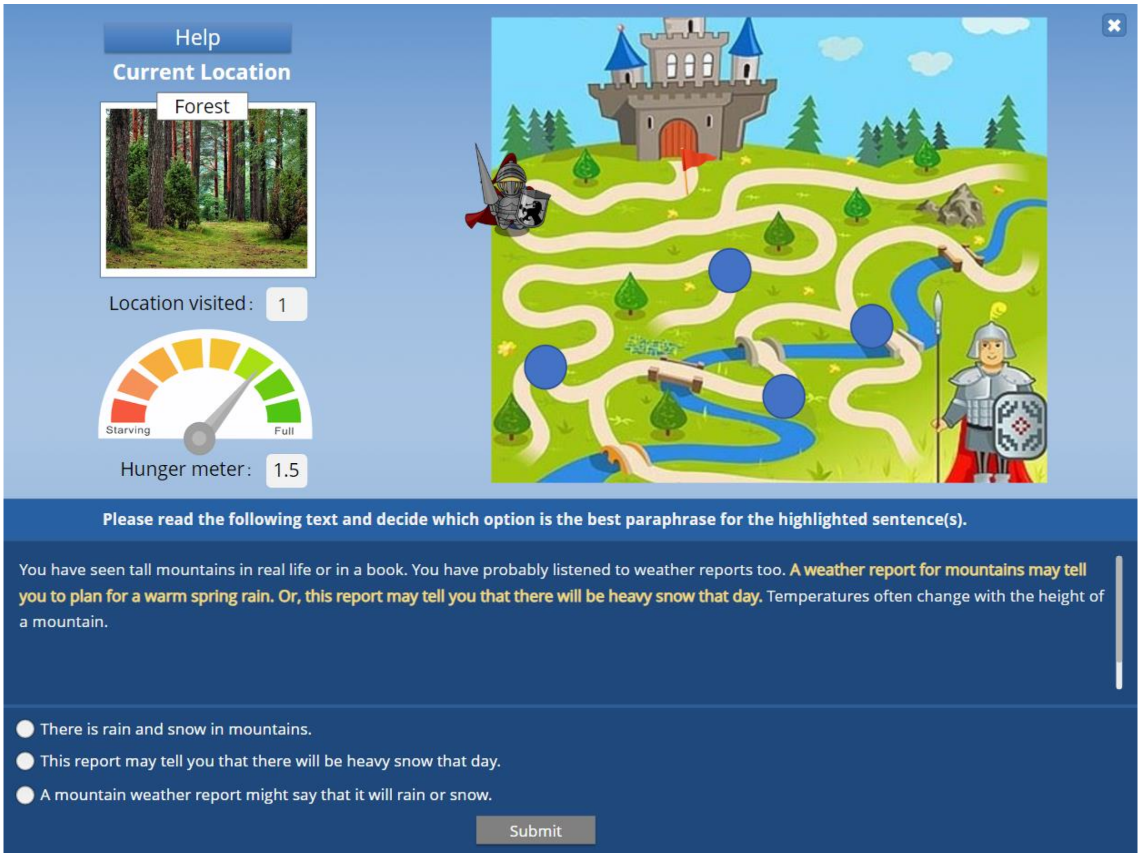
Which connective phrase links the sentences above? Fill in the blank by selecting a connecting flight.

Pick an answer, to board a flight

**Current Location :** Charlotte, NC

***Figure 2.*** CON-Artist game. This is an example question asking participants to choose the correct connecting phrase.

***Paraphrase Quest.*** Paraphrase Quest is included as a measure of how well participants can understand and rephrase individual sentences. Participants read short texts and are asked to select the appropriate paraphrase of the bolded sentence out of three options. Providing the correct paraphrase allows them to move forward on the map. The dependent variables are the proportion of correct responses and the average reaction time.



Location visited: 1

Hunger meter: 1.5

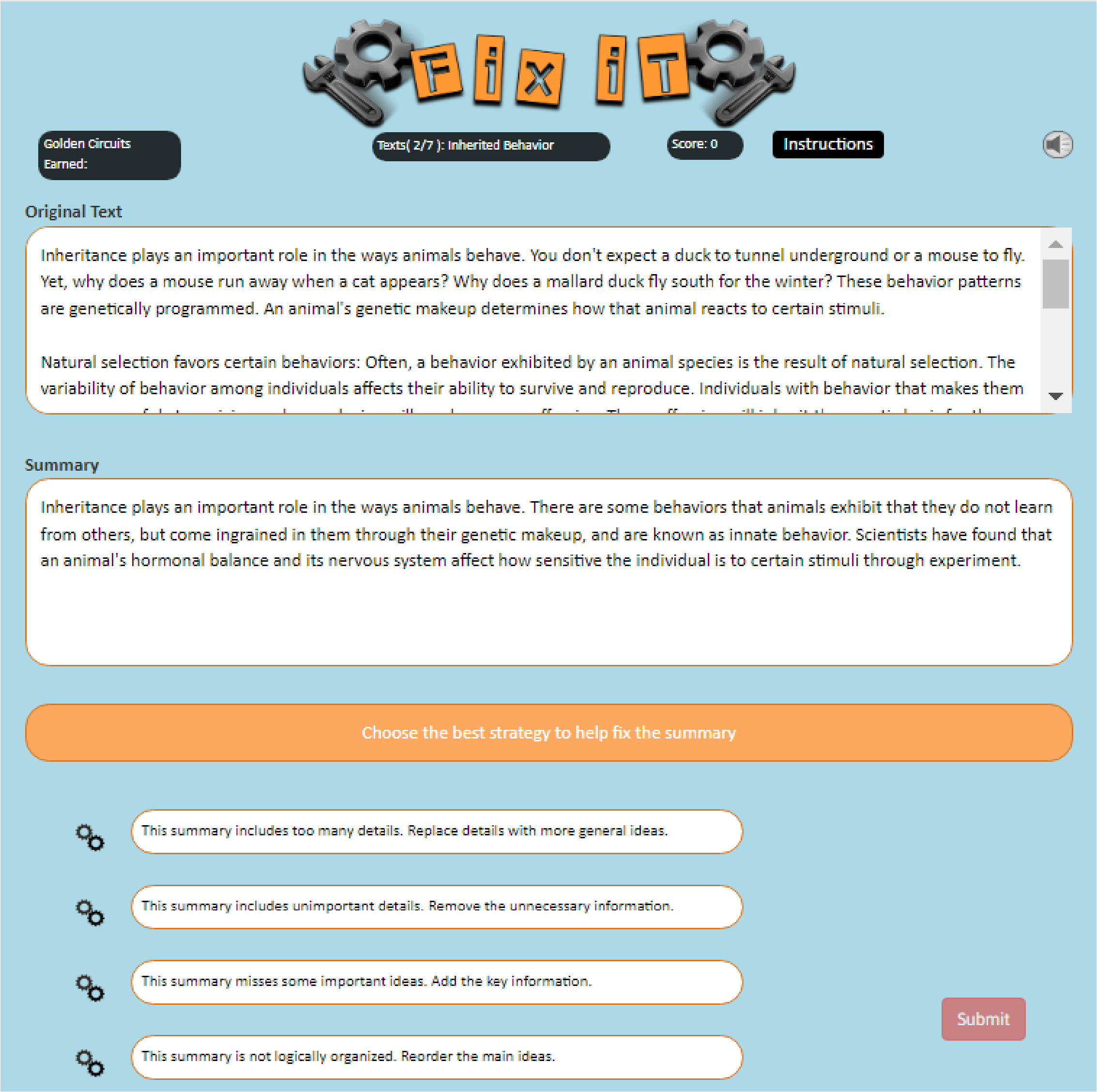
Please read the following text and decide which option is the best paraphrase for the highlighted sentence(s).

You have seen tall mountains in real life or in a book. You have probably listened to weather reports too. A weather report for mountains may tell you to plan for a warm spring rain. Or, this report may tell you that there will be heavy snow that day. Temperatures often change with the height of a mountain.

○ There is rain and snow in mountains.
○ This report may tell you that there will be heavy snow that day.
○ A mountain weather report might say that it will rain or snow.

***Figure 3.*** Paraphrase Quest game. Displayed is a sample question and the map participants see.

***Fix It.*** This task is included to measure how well participants can understand and summarize the main ideas of a text. Participants are given a short text to read, and a summary of the text below it. The texts are challenging science or history texts that are drawn from iSTART. Participants are asked to pick the answer choice out of four that best represents what is wrong with the summary, such as including too many details or missing important ideas. Participants receive a circuit if they respond correctly. The circuit is used to complete the circuit sudoku board at the end of the game. The dependent variables are the proportion of correct responses and the average reaction time.



Original Text

Inheritance plays an important role in the ways animals behave. You don't expect a duck to tunnel underground or a mouse to fly. Yet, why does a mouse run away when a cat appears? Why does a mallard duck fly south for the winter? These behavior patterns are genetically programmed. An animal's genetic makeup determines how that animal reacts to certain stimuli.

Natural selection favors certain behaviors: Often, a behavior exhibited by an animal species is the result of natural selection. The variability of behavior among individuals affects their ability to survive and reproduce. Individuals with behavior that makes them

Summary

Inheritance plays an important role in the ways animals behave. There are some behaviors that animals exhibit that they do not learn from others, but come ingrained in them through their genetic makeup, and are known as innate behavior. Scientists have found that an animal's hormonal balance and its nervous system affect how sensitive the individual is to certain stimuli through experiment.

Choose the best strategy to help fix the summary

This summary includes too many details. Replace details with more general ideas.

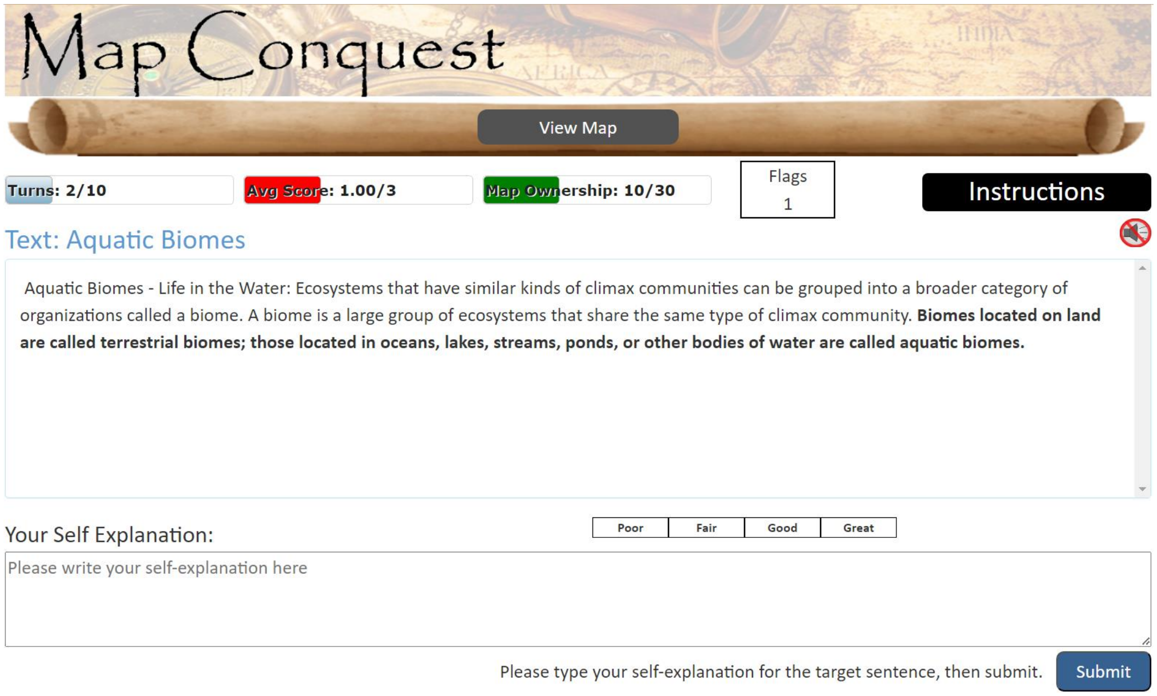
This summary includes unimportant details. Remove the unnecessary information.

This summary misses some important ideas. Add the key information.

This summary is not logically organized. Reorder the main ideas.

***Figure 4.*** Fix It game. This is the screen participants see while completing one of the questions.

***Map Conquest.*** Map Conquest is included to measure how well participants can explain what they read, and to compare the predictive value of this generative game to forced-choice games. Participants are asked to provide self-explanation for bold sentences in a text called “Aquatic Biomes”. The better the self-explanation is, the more flags the participants earn. Participants can strategically place the flags on the map to conquer territories in a mini game. The more flags they have, the easier it is to win.



Turns: 2/10

Avg Score: 1.00/3

Map Ownership: 10/30

Flags 1

Aquatic Biomes - Life in the Water: Ecosystems that have similar kinds of climax communities can be grouped into a broader category of organizations called a biome. A biome is a large group of ecosystems that share the same type of climax community. **Biomes located on land are called terrestrial biomes; those located in oceans, lakes, streams, ponds, or other bodies of water are called aquatic biomes.**

Please type your self-explanation for the target sentence, then submit.

***Figure 5.*** Map Conquest game.

## Natural Language Processing

**Procedure for evaluating self-explanations.** Self-explanations produced via the self-explanation exercise and Map Conquest were analyzed using the natural language processing tools TAACO (Crossley, Kyle, & Dascalu, 2019; Crossley et al., 2016) and Coh-Metrix (McNamara et al., 2014). TAACO measures local and global text cohesion. Coh-Metrix includes cohesion indices, but also more traditional text indices such as average word length, word frequency, word concreteness, and Flesch-Kincaid grade level. The self-explanations for a single participant on a single text were placed within separate paragraphs in the same text file.

**Selected indices.** Eight indices have been pre-selected from the many available indices to reduce the risk of Type I errors (Allen et al., 2021): word count, mean sentence length, lexical diversity (MTLD), ratio of causal particles to causal verbs, givenness, verb overlap, word2vec mean overlap between adjacent paragraphs, and word2vec similarity score between the self-explanation and the source text. These indices are used as the dependent variables for the self-explanation exercise and Map Conquest. Word count and mean sentence length were chosen as basic indicators of the overall quality and effort put into the self-explanation by the participant.

Lexical diversity refers to the extent to which unique words are used in text, rather than repeating words. Students who use more diverse words in their self-explanations tend to have better performance on text comprehension measures (Allen, Snow, & McNamara, 2015; Varner et al., 2013). When students are given self-explanation training, they tend to have higher lexical diversity in their self-explanations after training than before training (Allen et al., 2016).

The ratio of causal particles to causal verbs, also known as the causal ratio, is a measure of causal cohesion. Participants trained to self-explain tend to increase their causal ratio in their self-explanations (Allen et al., 2016), and students who are self-explaining tend to have a higher causal ratio than those who think-aloud in response to it (Allen, McNamara, & McCrudden, 2015). Givenness refers to the extent to which semantic information in a sentence can be understood from earlier parts of the self-explanation(s), and it tends to increase with self-explanation practice (Allen et al., 2016). Verb overlap refers to the extent to which verbs in the self-explanations contain similar meanings; participants tend to have more verb overlap in self- explanations than in think-alouds (Allen, McNamara, & McCrudden, 2015).

Word2vec (Mikolov et al., 2013) can be used to estimate semantic similarity between adjacent paragraphs and between two texts. Each word or phrase is represented in a vector-space model with a specific number of dimensions; TAACO uses 300 dimensions. Word2vec is trained using a neural network model, wherein each word's vector location is computed based on which words surround it. Words that appear in similar contexts are assigned similar vector locations, whereas words that rarely appear together are assigned disparate vector locations. Word2vec in TAACO was trained on magazines from the Corpus of Contemporary American English. One measure is the word2vec mean overlap between adjacent paragraphs; in this case, between adjacent self-explanations a participant provided in response to a text. Higher mean overlap indicates participants connecting information they referenced previously to new sections of the text. Mean overlap between adjacent paragraphs is higher for students who have high reading skill (Allen, Snow, & McNamara, 2015). Using the word2vec similarity score between the self-explanation and the source text, known as source-text similarity, can measure the extent to which the semantic meanings contained in a participant's self-explanation overlap with the semantic content in the text. A high similarity score indicates strong content overlap, signifying that the participant is paraphrasing the text. A low similarity score indicates the participant is elaborating on the text or including irrelevant information. Higher quality summaries tend to have higher source-text similarity (Crossley, Kim, et al., 2019).

### Procedure

The study was conducted through Qualtrics. Participants completed the study in a single session, in the presence of research assistants and/or graduate students. Participation occurred in person in an auditorium setting to increase participant recruitment and data quality. Participants first completed a consent form and demographics questions. They next completed the GMRT Reading Comprehension and Vocabulary sections, the Technology Acceptance Model Questionnaire, and the Cognitive Test Anxiety Scale. Then, they completed the Self-Explanation Task, followed by five games. Students were randomly assigned to one of two game orders. Order A was CON-Artist, Paraphrase Quest, Fix It, Map Conquest, and Vocab Flash. This order was randomly generated (see Table 1). Order B was the reverse order: Vocab Flash, Map Conquest, Fix It, Paraphrase Quest, and CON-Artist. Finally, participants completed the User Experience Survey and the Cognitive Test Anxiety Scale. Students also completed short enjoyment questionnaires after both sections of the Gates-MacGinitie Reading Test, the Self- Explanation Task, and each of the games. The session length was 2 hours.

Table 1

***Sample Experimental Procedure***

| Task | Estimated Time |
|---|---|
| Consent Form | 5 minutes |
| Demographics Questions | 5 minutes |
| Gates-MacGinitie Reading Comprehension & Enjoyment Survey | 25 minutes |
| Gates-MacGinitie Vocabulary & Enjoyment Survey | 10 minutes |
| Technology Acceptance Model Questionnaire | 5 minutes |
| Cognitive Test Anxiety Scale | 10 minutes |
| Self-Explanation Task & Enjoyment Survey | 15 minutes |
| Con-Artist & Enjoyment Survey | 5 minutes |
| Paraphrase Quest & Enjoyment Survey | 5 minutes |
| Fix It & Enjoyment Survey | 5 minutes |
| Map Conquest & Enjoyment Survey | 15 minutes |
| Vocab Flash & Enjoyment Survey | 5 minutes |
| User Experience Survey | 5 minutes |

# CHAPTER 5

# RESULTS

## Preliminary Analyses

Initial descriptives and correlations between variables are in Table 2 and Figure 6. For Map Conquest, the average of the game scores for each self-explanation, designed using natural language processing, were used rather than individual indices, to better illustrate comparisons across games. Split-half reliability was calculated using the Spearman-Brown formula. Reading Comprehension, Vocabulary, Cognitive Text Anxiety, Map Conquest, and Vocab Flash have good reliability, while Paraphrase Quest and Fix It have poor reliability. The Reading Comprehension and Vocabulary sections of the GMRT showed a positive, moderate correlation with Vocab Flash performance and positive, weak correlations with Paraphrase Quest, Fix It, and Map Conquest performance. Reading Comprehension and Vocabulary showed a negative, weak correlation with Cognitive Test Anxiety. All games showed positive, weak correlations with each other, except for Map Conquest, which showed a positive, weak correlation with Vocab Flash.

Table 2

***Descriptive Statistics and Correlations of Measures***

| Variable | *M* | *SD* | Observed Range | Split-Half | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. Reading Comprehension | 27.5 | 8.83 | 3 – 47 | .93 | — | | | | | |
| 2. Vocabulary | 28.72 | 8.25 | 8 – 44 | .92 | .40** | — | | | | |
| 3. Cognitive Test Anxiety | 69.72 | 16.67 | 31 – 107 | .94 | -.25** | -.11* | — | | | |
| 4. Paraphrase Quest | 0.67 | 0.23 | 0 – 1 | .41 | .19** | .13** | -.07 | — | | |
| 5. Fix It | 0.49 | 0.12 | .19 – 1 | .14 | .11* | .11* | .01 | .21** | — | |
| 6. Map Conquest | 1.61 | 0.61 | 0 – 3 | .89 | .22** | .15** | -.05 | .07 | .02 | — |
| 7. Vocab Flash | 0.63 | 0.15 | .09 – .98 | .80 | .56** | .46** | -.28** | .18** | .11* | .20** |

*Note*. *M* = mean. *SD* = standard deviation. * *p* < .05. ** *p* < .01. *n* = 405.

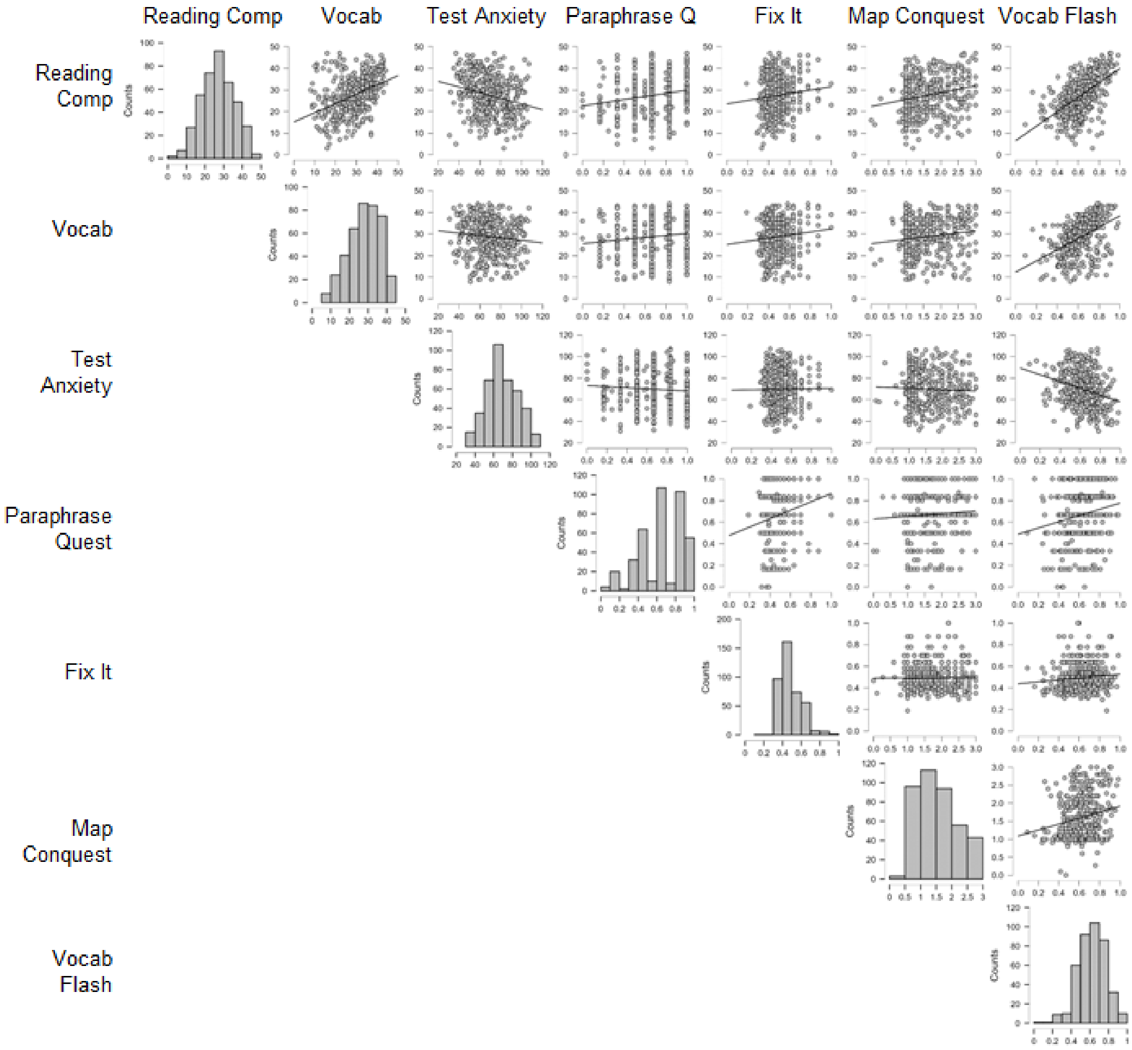


***Figure 6.*** Descriptives and correlations of reading comprehension, vocabulary, test anxiety, and game performance.

An independent t-test was conducted for each game to determine whether performance differed by order of completion (see Figure 7). Participants in Order A (n = 207) played in the order of CON-Artist, Paraphrase Quest, Fix It, Map Conquest, and finally Vocab Flash; Order B (n = 198) played the reverse order. Participants in Order A scored higher on Paraphrase Quest (t = 3.319, *p* < .001), which was the second game after CON-Artist (which was removed from analyses). Participants in Order B scored higher on Map Conquest (t = -6.995, *p* < .001) and Vocab Flash (t = -2.019, *p* = .044), which were the first two game that participants played. These results point toward the importance of limiting the number of games that participants play in any one session. All analyses were conducted both combined and separately for each order condition, given these differences by order condition.

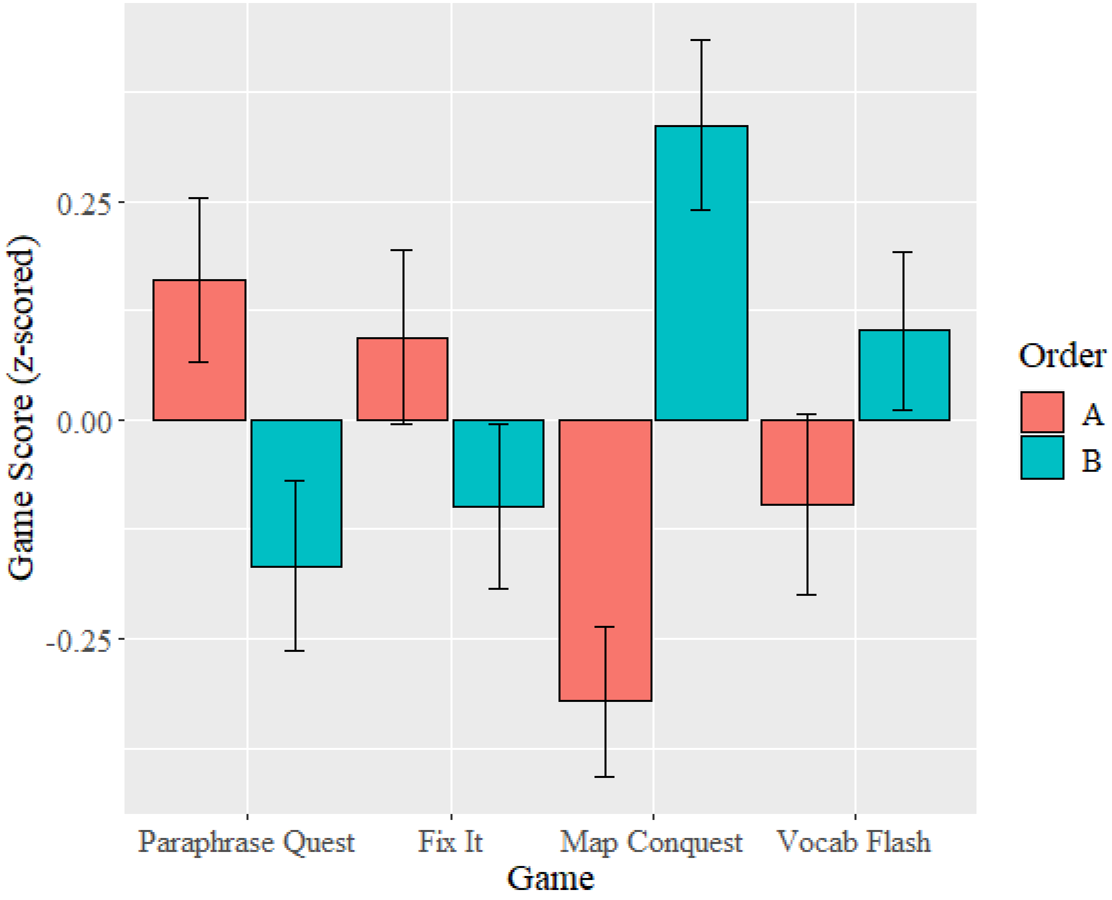


***Figure 7.*** Game performance by order condition. Order A indicates students who played in the order of Paraphrase Quest, Fix It, Map Conquest, and Vocab Flash. Order B indicates students who played in the order of Vocab Flash, Map Conquest, Fix It, and Paraphrase Quest.

## Primary Analyses

**Analysis 1a.** Does each game account for unique variance in GMRT Reading Comprehension performance? A hierarchical linear regression analysis was conducted using the “lm” function in R. Model 1 included accuracy in Vocab Flash. This measure reflects vocabulary knowledge; it was included in the first model because a prior study found that Vocab Flash accounted for 74% of variance (Fang et al., 2021). Model 2 added accuracy in Paraphrase Quest and Fix It, measures which reflect the ability to summarize at the sentence level and at the text level, respectively. Model 3 added the 8 selected natural language processing indices for Map Conquest (word count, sentence length, lexical diversity, causal ratio, givenness, paragraph overlap, verb overlap, and source similarity). Map Conquest was included last to demonstrate whether natural language processing of self-explanations provided additional information than could be obtained from identification games.

Cook’s distance was used to check for influential observations within this and all regression analyses. No distances greater than 1 were found, and so all participants were retained in the analyses. Model 1 accounted for approximately 31% of variance in reading comprehension, $F(1, 403) = 179.4$, $p < .001$ (see Figure 8). Models 2 and 3 did not account for variance beyond Model 1. No differences were found when conducting the analysis separately for participants in order A. When the analysis was conducted only for participants in game order B, Model 3 accounted for more variance than Model 1, F = 2.293, $p = 0.023$. The Map Conquest natural language indices representing source similarity (t = 2.835, $p = .005$) and givenness (t = 1.993, $p = .048$) predicted reading comprehension. Model 3 overall accounted for approximately 34% of variance in reading comprehension, F(11, 186) = 10.18, *p* < .001. All games were correlated with reading comprehension, but it appears that Vocab Flash alone is sufficient to explain most of the predicted variance in reading comprehension. A generative game such as Map Conquest might provide additional information, but only when participants have not been fatigued by playing several other games.

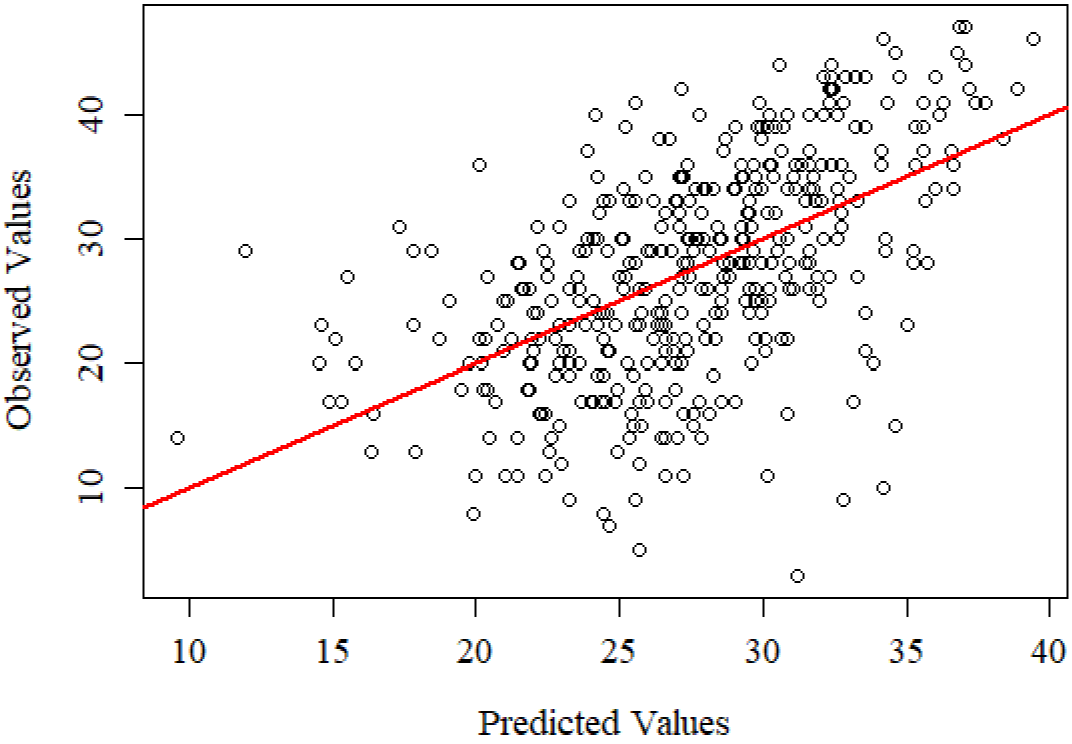


***Figure 8.*** Scatterplot of Model 1's predicted values for GMRT Reading Comprehension, compared to participants' obtained values for GMRT Reading Comprehension.

**Analysis 1b.** Does each game account for unique variance in GMRT Vocabulary performance? A hierarchical linear regression analysis was conducted. Model 1 included accuracy in Vocab Flash. Model 2 added accuracy in Paraphrase Quest and Fix It. Model 3 added the 8 selected natural language processing indices for Map Conquest (word count, sentence length, lexical diversity, causal ratio, givenness, paragraph overlap, verb overlap, and source similarity).

Model 1 accounted for approximately 21% of variance in vocabulary, *F*(1, 403) = 109.7, *p* < .001 (see Figure 9). Models 2 and 3 did not account for variance beyond Model 1. Results did not differ when examining each game order condition separately. This indicates that a simple vocabulary game may be sufficient for stealth assessment of vocabulary knowledge, without the need for generative or other games.

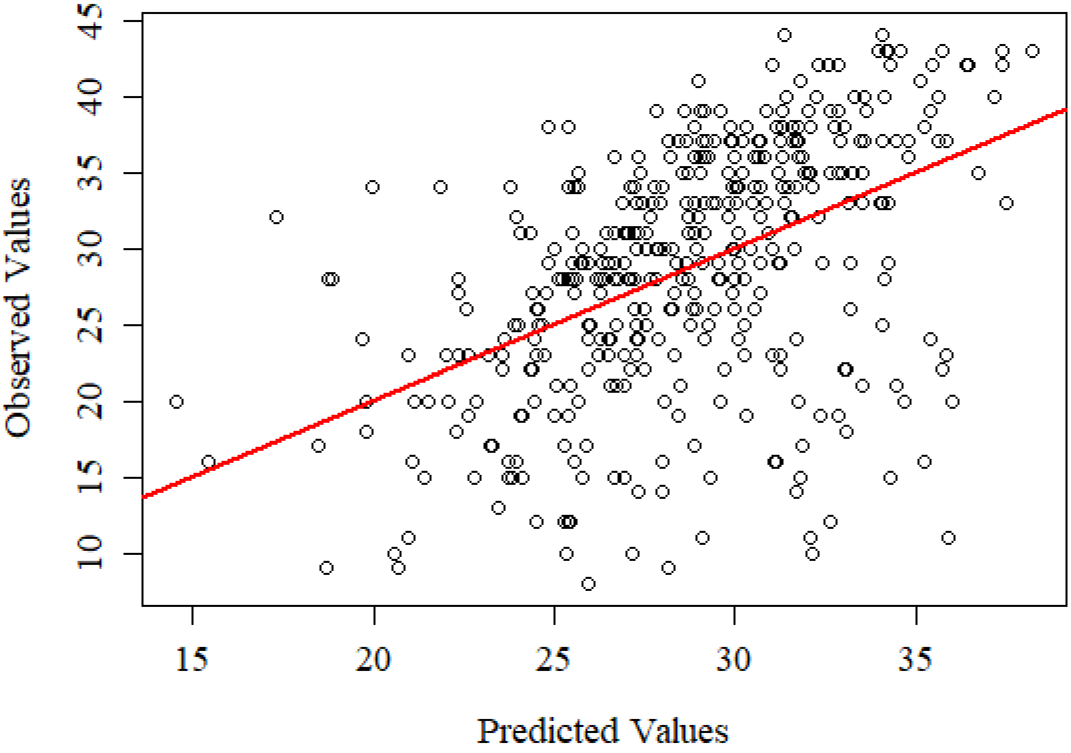


***Figure 9.*** Scatterplot of Model 1's predicted values for GMRT Vocabulary, compared to participants' obtained values for GMRT Vocabulary.

## Secondary Analyses

**Analysis 2a.** Is a person's enjoyment of literacy games predictive of literacy test performance above and beyond game performance? Two hierarchical linear regression analyses were conducted, one with GMRT reading comprehension as the dependent variable, and the other with GMRT vocabulary as the dependent variable. In both regressions, Model 1 included accuracy in Vocab Flash. Model 2 added accuracy in Paraphrase Quest and Fix It. Model 3 added the 8 selected natural language processing indices for Map Conquest (word count, sentence length, lexical diversity, causal ratio, givenness, paragraph overlap, verb overlap, and source similarity). Model 4 added enjoyment for each game, as well as the interaction between game enjoyment and game performance.

Model 4 predicted variance in reading comprehension beyond Models 1, 2, and 3, F = 2.586, *p* = .007. In the final model, Paraphrase Quest (t = 2.435, *p* = .015), the interaction between Paraphrase Quest score and Paraphrase Quest enjoyment (t = -2.063, *p* = .040), and the interaction between Vocab Flash score and Vocab Flash enjoyment (t = 2.757, *p* = .006) predicted reading comprehension (see Figures 10 and 11). The model predicted approximately 34% of variance in reading comprehension, F(20, 384) = 11.41, *p* < .001. Game performance helped predict reading comprehension for students who did not enjoy Paraphrase Quest. However, as student enjoyment of Paraphrase Quest increased, model accuracy in predicting reading comprehension test performance decreased. For these students, either Paraphrase Quest or the reading comprehension test are not accurate predictors of their true comprehension performance. Given the low reliability of Paraphrase Quest, likely Paraphrase Quest is a poor measure for these students. Meanwhile, among students who did poorly on Vocab Flash, those who did not enjoy it tended to have somewhat higher reading comprehension performance than those who did enjoy Vocab Flash. Among students who did well on Vocab Flash, those who enjoyed Vocab Flash tended to have somewhat higher reading comprehension performance than those who did not enjoy Vocab Flash. It appears that when Vocab Flash enjoyment aligned with Vocab Flash performance, reading comprehension performance was higher.

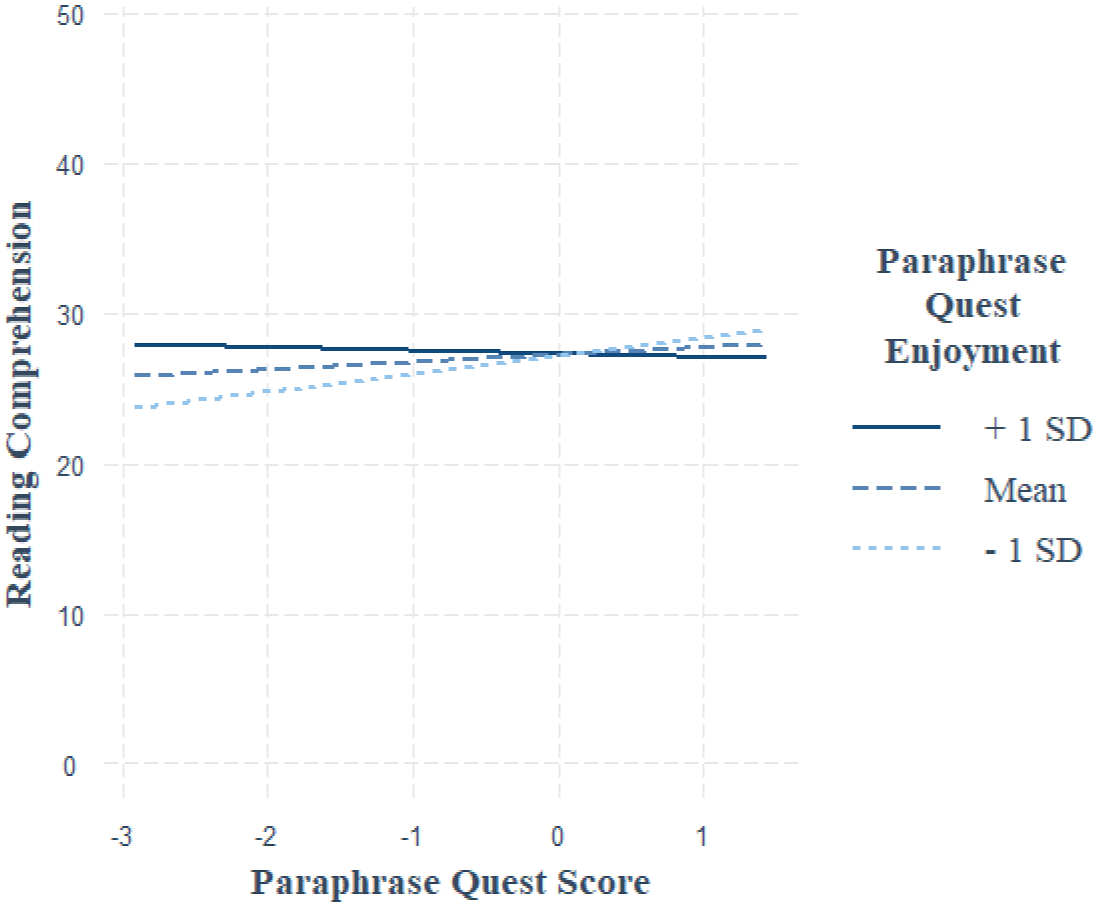


*Figure 10.* Illustration of how Paraphrase Quest enjoyment moderates the relationship between Paraphrase Quest and Reading Comprehension performance.

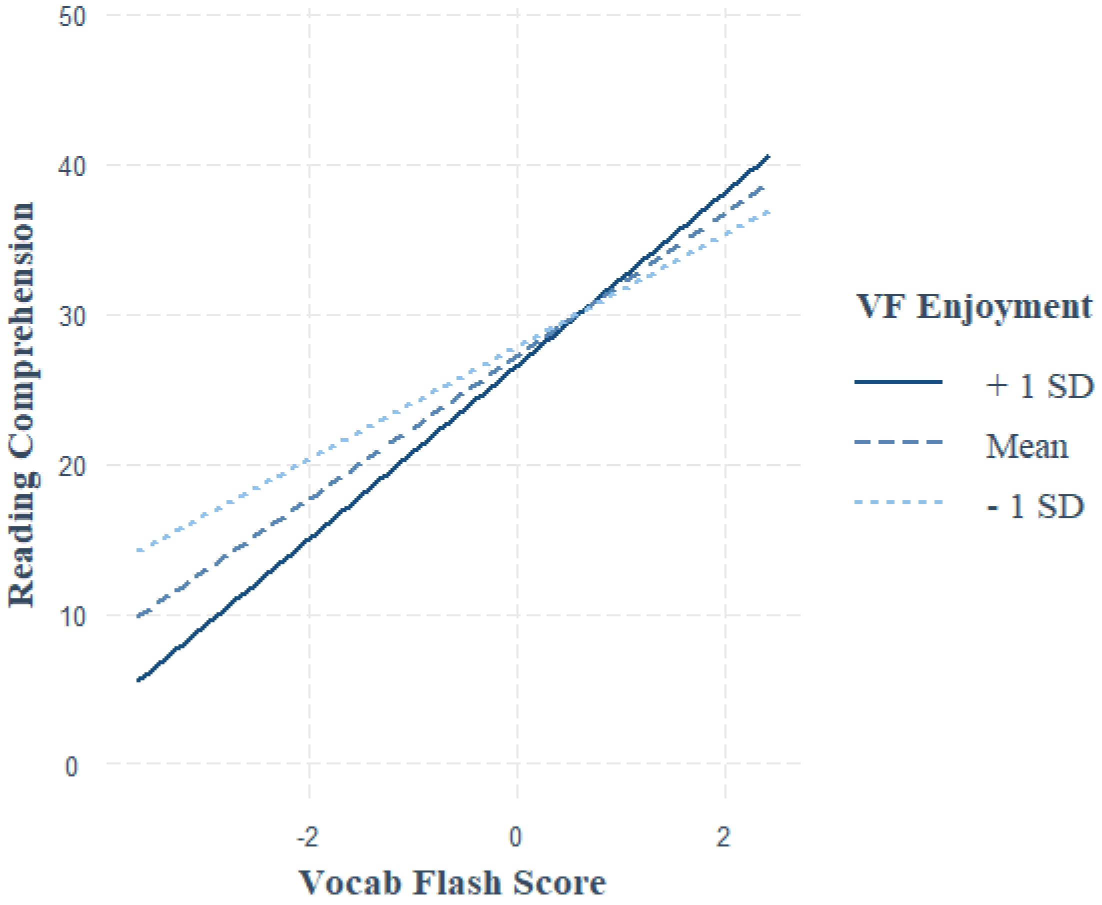


***Figure 11.*** Illustration of how Vocab Flash enjoyment moderates the relationship between Vocab Flash and Reading Comprehension performance.

Model 4 did not predict variance in vocabulary beyond Models 1, 2, and 3. When conducting analyses separately for Order A, Model 4 did not predict variance in reading comprehension or vocabulary beyond Models 1, 2, and 3. When conducting analyses separately for Order B, Model 4 predicted additional variance in GMRT Reading Comprehension but not GMRT Vocabulary. Map Conquest source similarity (t = 3.363, *p* < .001), Map Conquest mean number of words per self-explanation (t = -2.353, *p* = .020), and Vocab Flash enjoyment (t = -2.619, *p* = .010) predicted reading comprehension, and the model overall predicted about 38% of variance.

**Analysis 2b.** Is Map Conquest as predictive of reading test performance as the Self- Explanation Task? Two multiple linear regression analyses were conducted to analyze this question in terms of GMRT reading comprehension. In one, the Map Conquest indices were used to predict reading comprehension. In the other, the Self-Explanation Task indices were used to predict reading comprehension. The $R^2$ values were compared. Two multiple linear regression analyses were be conducted to analyze this question in terms of GMRT Vocabulary performance. In one, the Map Conquest indices were used to predict vocabulary. In the other, the Self-Explanation Task indices were used to predict vocabulary. The R^2 values were compared.

Map Conquest natural language indices predicted approximately 2.5% of variance in GMRT Reading Comprehension, F(8, 396) = 2.302, *p* = .020. The Self-Explanation Task natural language indices predicted approximately 8.1% of variance in reading comprehension, F(8, 396) = 5.45, *p* < .001. Map Conquest natural language indices did not predict vocabulary performance. Self-Explanation Task natural language indices predicted vocabulary, F(8, 396) = 2.298, *p* = .020. accounting for approximately 2.5% of variance; the Self-Explanation Task was not significantly more predictive than Map Conquest. This suggests that the Self-Explanation Task was a more useful indicator of reading comprehension than Map Conquest, indicating that fatigue and/or game elements may have detracted from the predictive validity of Map Conquest.

When the analysis was conducted using only participants from game order A, Map Conquest natural language indices predicted about 5.4% of variance in GMRT Reading Comprehension, F(8, 198) = 2.465, *p* = .014. The Self-Explanation Task natural language indices predicted approximately 10% of variance in reading comprehension, F(8, 198) = 3.885, *p* < .001. Map Conquest natural language indices predicted about 4.4% of variance in vocabulary, F(8, 198) = 2.191, *p* = .030. The Self-Explanation Task natural language indices did not predict vocabulary. Enjoyment and fatigue may play a role in the predictive validity of these tasks.

When the analysis was conducted using only participants from game order B, Map Conquest indices predicted 5.6% of variance in GMRT Reading Comprehension, F(8, 189) = 2.478, *p* = .014, while the Self-Explanation Task indices predicted 7.9% of variance in reading comprehension, F(8, 189) = 3.125, *p* = .002. Neither Map Conquest nor the Self-Explanation Task predicted GMRT Vocabulary, likely due to the decreased power of the analysis.

**Analysis 2c.** Do students enjoy literacy games more than non-game literacy assessments? A one-way within-subjects ANOVA was conducted to evaluate whether enjoyment differed across measures. Enjoyment differed between tasks, F(6, 2424) = 69.28, *p* < .001.

Because the omnibus test is significant, post-hoc tests were conducted to determine which measures differed from each other. The tasks, in order from least to most enjoyable, were the Self-Explanation Task, Map Conquest, the GMRT Reading Comprehension section, Fix It, Vocab Flash, Paraphrase Quest, and the GMRT Vocabulary section (see Figure 12). Fix It, Vocab Flash, and Paraphrase Quest were all more enjoyable than the Self-Explanation Task. Vocab Flash and Paraphrase Quest were also more enjoyable than the GMRT Reading Comprehension Section. The results did not differ when examining each game order separately. Overall, the games showed the potential to be more enjoyable than a traditional reading test, but not more than a traditional vocabulary test.

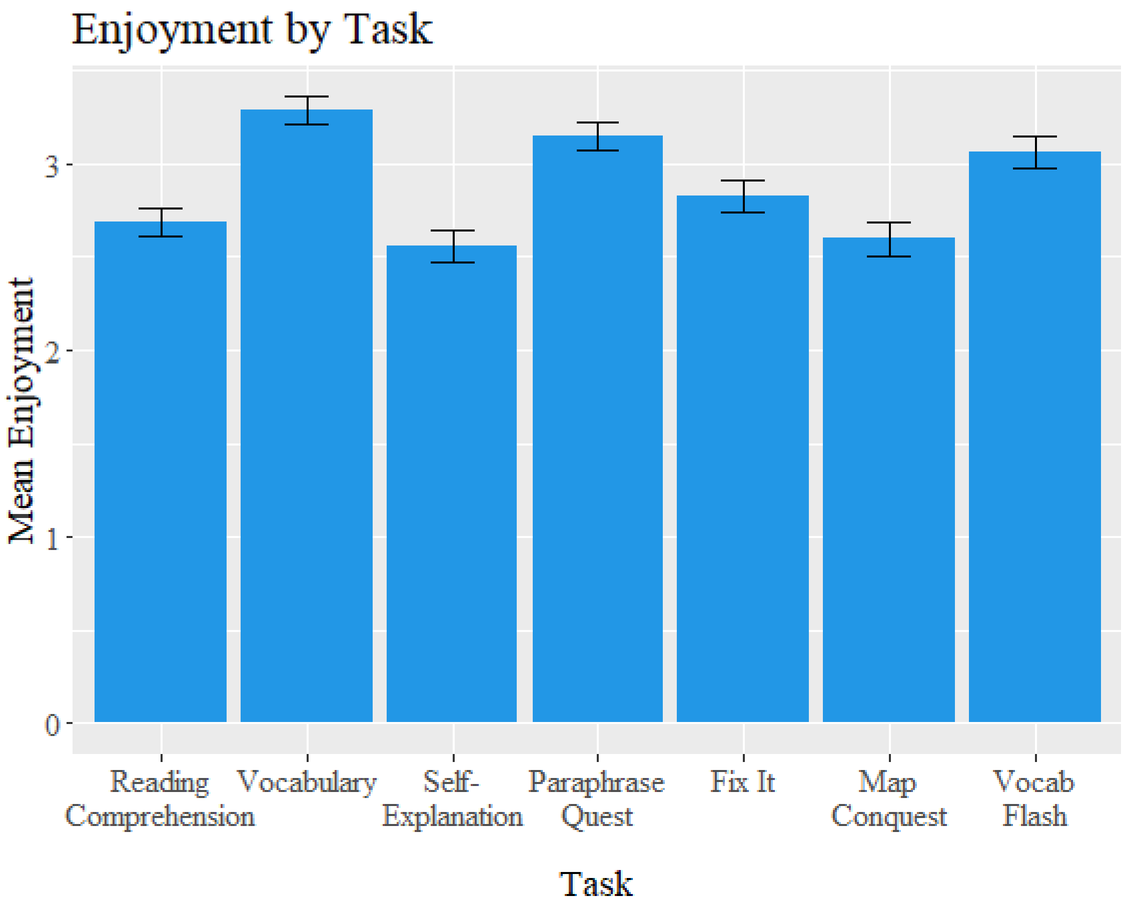


***Figure 12*****.** Reported mean enjoyment by task.

**Analysis 2d.** Is test anxiety, as measured by a test anxiety scale, more related to test performance than game performance? Cognitive test anxiety was correlated to GMRT Reading Comprehension, GMRT Vocabulary, and performance on each of the games. Test anxiety was only correlated with Vocab Flash (r = -0.28), reading comprehension (r = -0.25), and vocabulary (r = -0.11). Given that these were the only timed tasks in the study, it seems that test anxiety relates to performance in timed situations, but not untimed ones, regardless of the type of task.

# CHAPTER 6

# DISCUSSION

The primary goal of the current study was to examine the extent to which games can be used as stealth assessments of literacy skills. Five literacy games, representing different tasks related to reading comprehension, were selected: Vocab Flash, CON-Artist, Paraphrase Quest, Fix It, and Map Conquest. Students' performance on the games was used to predict reading comprehension skill and vocabulary knowledge, as measured by the Gates-MacGinitie Reading Test (GMRT). Vocab Flash predicted both reading comprehension skill and vocabulary knowledge, indicating that a simple vocabulary game can be used as stealth assessment of literacy skills. The success of game-based vocabulary assessment in predicting reading comprehension skill is not a surprise, because reading comprehension skill is strongly correlated to vocabulary knowledge (Cain & Oakhill, 2014). Vocab Flash is shorter and more enjoyable than the GMRT Reading Comprehension section, and it can also be used as a learning activity. Games like Vocab Flash could therefore be used in classrooms to give students and teachers quick assessments of reading skills while increasing student learning.

Vocab Flash was the only game that predicted reading comprehension skill and vocabulary knowledge without considering enjoyment and game order. Therefore, the most precise assessments of reading comprehension skill and vocabulary knowledge should include a vocabulary component. Even if reading comprehension skill and vocabulary knowledge were predicted without using Vocab Flash, only Paraphrase Quest was a significant predictor, and it could not account for as much variance as Vocab Flash. Paraphrase Quest and Fix It may have been limited as predictors because of their low reliability, perhaps due to the low number of questions used. Additionally, Vocab Flash was the only game that adapted question difficulty to the student's performance. Future research should consider incorporating adaptivity into other types of literacy games, as this may improve the precision of the games' estimates of literacy skills. A long assessment, such as the GMRT, includes questions with variable difficulty levels, so that the student's skills are precisely assessed. Game-based stealth assessment may benefit from incorporating adaptivity to successfully estimate students' skills within a shorter timeframe. Adaptivity also benefits the student by creating an optimum level of difficulty for learning and enjoyment.

Secondary analyses involving enjoyment indicate that there is potential to using non-vocabulary games in assessing literacy skills, but also a need for further research to make the games precise indicators for all students. Paraphrase Quest was a significant predictor when interactions between game enjoyment and game performance were included. For both Paraphrase Quest and Vocab Flash, enjoyment moderated the relationship between game performance and test performance. Paraphrase Quest was less predictive for students who enjoyed the game, perhaps these students were distracted by the game components. On the other hand, Vocab Flash was less predictive for students who did not enjoy the game, particularly for students who played the game last. Students with low enjoyment of Vocab Flash may not have displayed their full vocabulary knowledge, impacting the validity of the measurement. One possible solution is to improve the game elements to ensure the games are enjoyable to all students, while complementing instead of distracting from the game's literacy component. The study's open-ended feedback suggests students would prefer the game elements to be more closely related to the texts they read and to be more interactive. For example, Paraphrase Quest could have texts related to medieval Europe to match the knight graphics. It could also allow students to move the knight themselves, instead of the game automatically moving the knight. Such changes might give the games a better chance of engaging all students and limit the effects of enjoyment on game performance.

Similarly, the results suggest that Map Conquest could have been a more precise predictor of reading comprehension skills if the game elements were clearer and study fatigue was reduced. Map Conquest was less predictive than an equivalent self-explanation task. In open-ended feedback, many students noted that they found the game instructions confusing, impacting their ability to successfully play the game. Implementing clear instructions would help more students use their full abilities while playing the game. The predictive validity of Map Conquest was further impacted by the length of the study session; Map Conquest was predictive for students in game order B, but not game order A. Two hours of study participation may be too long for demanding literacy tasks; future studies should limit the participation time to 90 minutes.

Because a simple vocabulary game can predict student performance in standardized tests, games are plausibly an additional method for teachers to monitor student progress. However, special care should be taken so that the game mechanics do not impair the validity of the stealth assessment. Many traditional assessments have been carefully developed over many years to have reliability and validity (Schrank & Wendling, 2018). Thus, a mix of stealth assessments and traditional assessments may be ideal in a classroom setting. Such a combination would help students and teachers more accurately assess students' learning games, increase student enjoyment and engagement, and minimize classroom time spent administering tests.

# REFERENCES

Allen, L. K., Creer, S. D., & Poulos, M. C. (2021). Natural language processing as a technique for conducting text-based research. *Lang Linguist Compass*. doi:10.1111/lnc3.12433

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Cohesive features of deep text comprehension processes. In J. Trueswell, A. Papafragou, D. Grodner, & D. Mirman (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society in Philadelphia, PA* (pp. 2681-2686). Austin, TX: Cognitive Science Society. Retrieved from https://eric.ed.gov/?id=ED577140

Allen, L. K., McNamara, D. S., & McCrudden, M. (2015). Change your mind: Investigating the effects of self-explanation in the resolution of misconceptions. In D. C. Noelle, R. Dale, A. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 78-83). Pasadena, CA: Cognitive Science Society. Retrieved from https://cogsci.mindmodeling.org/2015/papers/0024/paper0024.pdf

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 *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge (LAK '15)* (pp. 246-254). New York, NY: Association for Computing Machinery. doi:10.1145/2723576.2723617

Anderson, R. C., & Nagy, W. (1992). The vocabulary conundrum. *American Educator, 16*. Retrieved from https://files.eric.ed.gov/fulltext/ED354489.pdf

Arum, R., & Roksa, J. (2011). Limited learning on college campuses. *Society, 48*(3), 203-207. doi:10.1007/s12115-011-9417-8

Barab, S. A., Pettyjohn, P., Gresalfi, M., Volk, C., & Solomou, M. (2012). Game-based curriculum and transformational play: Designing to meaningfully position person, content, and context. *Computers & Education, 58*, 518–533. doi:10.1016/j.compedu.2011.08.001

Boonthum-Denecke, C., McCarthy, P. M., Lamkin, T., Jackson, G. T., Magliano, J., & McNamara, D. S. (2011). Automatic natural language processing and the detection of reading skills and reading comprehension. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 234-239). Menlo Park, CA: AAAI Press. Retrieved from https://www.researchgate.net/publication/221437842_Automatic_Natural_Language_Processing_and_the_Detection_of_Reading_Skills_and_Reading_Comprehension

Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is vocabulary more important for some aspects of comprehension? *L'Annee Psychologique, 114*, 647-662. doi:10.4074/S0003503314004035

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270-295. doi:10.1006/ceps.2001.1094

Crossley, S. A., Kim, M., Allen, L., & McNamara, D. (2019). Automated summarization evaluation (ASE) using natural language processing tools. In *Artificial Intelligence in Education: 20th International Conference, Proceedings, Part I 20* (pp. 84-95). Chicago, IL: Springer International Publishing. doi:10.1007/978-3-030-23204-7_8

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavioral Research Methods, 51*, 14-27. doi:10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods 48*(4), 1227-1237. doi:10.3758/s13428-015-0651-7

Crossley, S. A., McNamara, D. S., Baker, R., Wang, Y., Paquette, L., Barnes, T., & Bergner, Y. (2015). Language to completion: Success in an educational data mining massive open online class. 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. 388-391). Madrid, Spain: Springer. Retrieved from https://www.educationaldatamining.org/EDM2015/proceedings/short388-391.pdf

Crosson, A.C., and Lesaux, N.K. (2011). Does knowledge of connectives play a unique role in the reading comprehension of English learners and English-only students?. *Journal of Research in Reading*, *36*, 241-260. doi:10.1111/j.1467-9817.2011.01501.x

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q., 13*(3), 319–339. doi:10.2307/249008

Fang, Y., Li, T., Roscoe, R. D., & McNamara, D. S. (2021). Predicting Literacy Skills via Stealth Assessment in a Simple Vocabulary Game. In R. A. Sottilare, & J. Schwarz (Eds.), *3rd International Conference on Adaptive Instructional Systems, Proceedings* (pp. 32-44). doi:10.1007/978-3-030-77873-6_3

*Gates-MacGinitie Reading Tests*. 1989. *Technical Report for Gates-MacGinitie Reading Tests Form S*. Chicago, Illinois: Riverside Publishing.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 2,* 371-398. doi:10.1111/j.1756-8765.2010.01081.x

Graham, S., & Herbert, M. A. (2010). Writing to read: Evidence for how writing can improve reading (a Carnegie Corporation Time to Act report). Washington, DC: Alliance of Exceptional Education. Retrieved from https://lincs.ed.gov/state-resources/federal-initiatives/teal/publications/writing-read

Hagaman, J. L., Casey, K. J., & Reid, R. (2016). *Preventing School Failure, 60*(1), 43-52. doi:10.1080/1045988X.2014.966802

Haynes, J. E., & Fillmer, H. T. (1984), Paraphrasing and reading comprehension. *Reading World, 24(1),* 76-79. doi:10.1080/19388078409557804

Kintsch, W. M. (1988). The Role of Knowledge in Discourse Comprehension: A Construction- Integration Model. *Psychological Review, 95*(2), 163–182. doi:10.1037/0033-295X.95.2.163

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

Klimmt, C., Blake, C., Hefner, D., Vorderer, P., & Roth, C. (2009). Player performance, satisfaction, and video game enjoyment. In *International conference on entertainment computing* (pp. 1-12). Berlin: Springer. doi:10.1007/978-3-642-04052-8_1

Kohnen, N., & Retelsdorf, J. (2019). The role of knowledge of connectives in comprehension of a German narrative text. *Journal of Research in Reading, 42,* 371-388. doi:10.1111/1467-9817.12273

Magliano, J. P., Millis, K. K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107-136). Mahwah, NJ: Lawrence Erlbaum Associates. Retrieved from https://www.researchgate.net/publication/237122752_A_Multidimensional_Framework_to_Evaluate_Reading_Assessment_Tools

McCarthy, K. S., Likens, A. D., Johnson, A. M., Guerrero, T. A., & McNamara, D. S. (2018). Metacognitive overload!: Positive and negative effects of metacognitive prompts in an intelligent tutoring system. *International Journal of Artificial Intelligence in Education, 28,* 420-438. doi:10.1007/s40593-018-0164-5

McNamara, D. S. (2009). The importance of teaching reading strategies. *Perspectives on Language and Literacy,* 34-40. The International Dyslexia Association.

McNamara, D. S. (2017). Self-explanation and reading strategy training (SERT) improves low-knowledge students’ science course performance. *Discourse Processes*, 1-14. doi:10.1080/0163853X.2015.1101328

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written communication, 27*(1), 57-86. doi:10.1177/0741088309351547

McNamara, D.S., Graesser, A.C., McCarthy, P.M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* New York, NY: Cambridge University Press.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. *Psychology of learning and motivation, 51*, 297-384. doi:10.1016/S0079-7421(09)51009-2

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*. Retrieved from https://arxiv.org/abs/1310.4546

Millis, K. K., & Just, M. A. (1994). The influence of connectives in sentence comprehension. *Journal of Memory and Language, 33*, 128-147. doi:10.1006/jmla.1994.1007

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives, 1*(1), 3-62. doi:10.1207/S15366359MEA0101_02

National Assessment of Educational Progress (2019). *NAEP report card: 2019 NAEP reading assessment*. U.S. Department of Education, Institute of Education Sciences. Retrieved from https://www.nationsreportcard.gov/highlights/reading/2019/

National Center for Education Statistics (2019). *Adult Literacy in the United States.* U.S. Department of Education, Institute of Education Sciences. Retrieved from https://nces.ed.gov/pubs2019/2019179/index.asp

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of educational psychology, 98*(3), 554. doi:10.1037/0022-0663.98.3.554

Ozuru, Y., Briner, S., Best, R., & McNamara, D. S. (2010). Contributions of self-explanation to comprehension of high-and low-cohesion texts. *Discourse Processes*, *47*(8), 641-667. doi:10.1080/01638531003628809

Ozuru, K., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing Comprehension Measured by Multiple-Choice and Open-Ended Questions. *Canadian Journal of Experimental Psychology, 67*(3), 215-227. doi:10.1037/a0032918

Parsons, J., & Taylor, L. (2011). Improving student engagement. *Current issues in education, 14*(1). Retrieved from https://eric.ed.gov/?id=EJ938960

Phillips, L. M., Norris, S. P., Osmond, W. C., & Maynard, A. M. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. *Journal of Educational Psychology, 94*(1), 3. doi:10.1037/0022-0663.94.1.3

Schrank, F. A., & Wendling, B. J. (2018). The Woodcock–Johnson IV: Tests of cognitive abilities, tests of oral language, tests of achievement. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 383–451). The Guilford Press.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction*, *55*(2), 503-524. Information Age Publishing. Retrieved from https://myweb.fsu.edu/vshute/pdf/shute%20pres_h.pdf.

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. doi: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. doi:10.1016/j.chb.2016.05.047

Sinclair, J., Jang, E. E., & Rudzicz, F. (2021). Using machine learning to predict children’s reading comprehension from linguistic features extracted from speech and writing. *Journal of Educational Psychology, 113*(6), 1088–1106. doi:10.1037/edu0000658

Smolkowski, K., & Cummings, K. D. (2016). Evaluation of the DIBELS (Sixth Edition) Diagnostic System for the Selection of Native and Proficient English Speakers at Risk of Reading Difficulties. *Journal of Psychoeducational Assessment, 34*(2), 103–118. doi:10.1177/0734282915589017

Snyder, L., Caccamise, D. & Wise, B. (2005). The Assessment of Reading Comprehension: Considerations and Cautions. *Topics in Language Disorders, 25*(1). doi:10.1097/00011363-200501000-00005

Taylor, R., O'Reilly, T., Sinclair, G., & McNamara, D. (2006). Enhancing learning of expository science texts in a remedial reading classroom via iSTART. In Barab, S. A., Hay, K. E., & Hickey, D. T. (Eds.), *Proceedings of The International Conference of the Learning Sciences, Volume 2* (pp. 765-770). Bloomington, Indiana, USA: International Society of the Learning Sciences. doi:10.22318/icls2006.765

Van den Broek, P. W., Helder, A., & Van Leijenhorst, L. (2012). Sensitivity to structural centrality: Developmental and individual differences in reading comprehension skills. In M. A. Britt, S. R. Goldman, & J.-F. Rouet (Eds.), *Reading: From words to multiple texts.* New York, NY: Routledge, Taylor & Francis Group.

Van Silfhout, G., Evers-Vermeul, J., Sanders, T. (2015). Connectives as processing signals: How students benefit in processing narrative and expository texts. *Discourse Processes, 52*, 47-76. doi:10.1080/0163853X.2014.905237

Varner, L. K., Jackson, G. T., Snow, E. L., & McNamara, D. S. (2013). Does Size Matter? Investigating User Input at a Larger Bandwidth. In *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference.* Association for the Advancement of Artificial Intelligence. Retrieved from https://eric.ed.gov/?id=ED588486

Volodina, A., Heppt, B., & Weinert, S. (2021). Relations between the comprehension of connectives and school performance in primary school. *Learning and Instruction, 74.* doi:10.1016/j.learninstruc.2020.101430

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. doi:10.1016/j.jad.2017.11.048

Wilson, K. A., Bedwell, W. L., Lazzara, E. H., Salas, E., Burke, S., Estock, J. L., Orvis, K. L., & Conkey, C. (2009). Relationships between game attributes and learning outcomes: Review and research proposals. *Simulation and Gaming, 40*, 217-266. doi:10.1177/1046878108321866

# APPENDIX A

# DEMOGRAPHICS QUESTIONS

1. Sex
    1. Male
    2. Female
    3. Intersex
    4. Prefer not to say
2. Gender
    1. Man
    2. Woman
    3. Non-binary / third gender
    4. Prefer not to say
3. Age
    1. _____
4. Ethnicity
    1. Hispanic or Latino
    2. American Indian or Alaska Native
    3. Asian
    4. Black or African American
    5. Native Hawaiian or Other Pacific Islander
    6. Caucasian or White
    7. Multiracial
    8. Other
    9. Prefer not to say
5. What is the highest degree of education you have completed?
    1. Less than high school degree
    2. High School Degree
    3. Associate’s Degree or Trade School
    4. Bachelor’s Degree

    5. Master’s or Doctorate Degree

6. Please select the option “Agree”
    1. Strongly disagree
    2. Disagree
    3. Neither agree nor disagree
    4. Agree
    5. Strongly agree
7. Is English your first language?
    1. Yes
    2. No
8. What is your native language?
    1. _____
9. How many years have you been studying English?
    1. Less than 1 year
    2. 1 year
    3. 2 years
    4. 3 years
    5. 4 years
    6. 5 years
    7. 6 years
    8. 7 or more years
10. What types of texts do you generally write in English? (You may select multiple responses.)
    1. Emails
    2. Letters
    3. Notes
    4. Essays

5. Research Papers
6. Reports
7. Creative Writing

11. Do you like writing in English?

1. I don’t like it at all
2. I don’t like it
3. I have no feelings about it
4. I like it
5. I like it a lot

# APPENDIX B

# TECHNOLOGY ACCEPTANCE MODEL QUESTIONNAIRE

Please read the following statements and rate how they apply to you.

| | Strongly Disagree | Disagree | Somewhat Disagree | Neither Agree nor Disagree | Somewhat Agree | Agree | Strongly Agree |
|---|---|---|---|---|---|---|---|
| Using computer systems improves my performance while learning. | o | o | o | o | o | o | o |
| Using computer systems while learning increases my productivity. | o | o | o | o | o | o | o |
| Using computer systems enhances my effectiveness while learning. | o | o | o | o | o | o | o |
| I find computer systems to be useful while learning. | o | o | o | o | o | o | o |
| My interactions with computer systems are clear and understandable. | o | o | o | o | o | o | o |
| Interacting with computer systems does not require a lot of my mental effort. | o | o | o | o | o | o | o |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| I find computer systems easy to use. | o | o | o | o | o | o | o |
| I have control over computer systems. | o | o | o | o | o | o | o |

# APPENDIX C

# COGNITIVE TEST ANXIETY SCALE

1. I lose sleep over worrying about examinations.
2. While taking an important examination, I find myself wondering whether the other students are doing better than I am.
3. I have less difficulty than the average college student in getting test instructions straight.
4. I tend to freeze up on things like intelligence tests and final exams.
5. I am less nervous about tests than the average college student.
6. During tests, I find myself thinking of the consequences of failing.
7. At the beginning of a test, I am so nervous that I often can't think straight.
8. The prospect of taking a test in one of my courses would not cause me to worry.
9. I am more calm in test situations than the average college student.
10. I have less difficulty than the average college student in learning assigned chapters in textbooks.
11. My mind goes blank when I am pressured for an answer on a test.
12. During tests, the thought frequently occurs to me that I may not be too bright.
13. I do well in speed tests in which there are time limits.
14. During a course examination, I get so nervous that I forget facts I really know.
15. After taking a test, I feel I could have done better than I actually did.
16. I worry more about doing well on tests than I should.
17. Before taking a test, I feel confident and relaxed.
18. While taking a test, I feel confident and relaxed.
19. During tests, I have the feeling that I am not doing well.
20. When I take a test that is difficult, I feel defeated before I even start.
21. Finding unexpected questions on a test causes me to feel challenged rather than panicky.
22. I am a poor test taker in the sense that my performance on a test does not show how much I really know about a topic.
23. I am not good at taking tests.
24. When I first get my copy of a test, it takes me a while to calm down to the point where I can begin to think straight.
25. I feel under a lot of pressure to get good grades on tests.
26. I do not perform well on tests.
27. When I take a test, my nervousness causes me to make careless errors.

# APPENDIX D

# SELF-EXPLANATION TASK

**Self-Explanation Exercise**

You will now be asked to read a few passages of text and provide your own self-explanations. When writing your explanation, some good strategies are paraphrasing, bridging, and elaboration. Paraphrasing is restating the sentence in your own words. Bridging is connecting the sentence to earlier sentences in the text. Elaboration is connecting the information in the sentence to what you already know about the topic. Below you'll see a passage and examples of each strategy.

Passage: "For as long as there have been forests, lightning has been igniting forest fires. In the past, these fires simply burned themselves out because there was no way to stop them. With the development of modern technology (airplanes, powerful water pumps, chemicals), we now have the means to put out many of these fires."

Example of paraphrase: "Forest fires can be snuffed out now with new equipment like strong water pumps."

Example of bridging: "We can put out fires that have been ignited by lightning, while before we could do nothing but watch."

Example of elaboration: "Since humans have more information now about the chemical structure of fire and engineering principles, we've been able to build new technologies that can effectively fight large fires."

You can see that there are multiple strategies that can be used to self-explain the passage. You don't have to use every strategy in each self-explanation; use whichever ones you find useful.

1. Please read the text and provide a self-explanation for the bolded sentence below.

Red blood cells have the vital role of carrying oxygen to all of the cells in the body. **They also pick up waste carbon dioxide for removal.**

______________________________________________________________

2. Please read the text and provide a self-explanation for the bolded sentence below.

Red blood cells have the vital role of carrying oxygen to all of the cells in the body. They also pick up waste carbon dioxide for removal. These cells are the most numerous of the blood cells. **The disk shape of red blood cells results in a large surface area, which enables them to be efficient at gas diffusion.**

______________________________________________________________

3. Please read the text and provide a self-explanation for the bolded sentence below.

Red blood cells have the vital role of carrying oxygen to all of the cells in the body. They also pick up waste carbon dioxide for removal. These cells are the most numerous of the blood cells. The disk shape of red blood cells results in a large surface area, which enables them to be efficient at gas diffusion.

Red blood cells contain a large, complex protein called hemoglobin. **Hemoglobin binds to the oxygen and carbon dioxide that the red blood cells transport.**

______________________________________________________________

4. Please read the text and provide a self-explanation for the bolded sentence below.

Red blood cells have the vital role of carrying oxygen to all of the cells in the body. They also pick up waste carbon dioxide for removal. These cells are the most numerous of the blood cells. The disk shape of red blood cells results in a large surface area, which enables them to be efficient at gas diffusion.

Red blood cells contain a large, complex protein called hemoglobin. Hemoglobin binds to the oxygen and carbon dioxide that the red blood cells transport. Each red blood cell contains about 250 million hemoglobin molecules, each carrying four molecules of oxygen. **Hemoglobin also contains iron, which gives blood its red color.**

______________________________________________________________

5. Please read the text and provide a self-explanation for the bolded sentence below.

Red blood cells have the vital role of carrying oxygen to all of the cells in the body. They also pick up waste carbon dioxide for removal. These cells are the most numerous of the blood cells. The disk shape of red blood cells results in a large surface area, which enables them to be efficient at gas diffusion.

Red blood cells contain a large, complex protein called hemoglobin. Hemoglobin binds to the oxygen and carbon dioxide that the red blood cells transport. Each red blood cell contains about 250 million hemoglobin molecules, each carrying four molecules of oxygen. Hemoglobin also contains iron, which gives blood its red color. Molecular oxygen can also be transported by another route, in dissolved blood plasma. However, oxygen is poorly soluble in water, so only about 1.5% is carried in dissolved form. **Therefore, most oxygen is carried by hemoglobin.**

______________________________________________________________

6. Please read the text and provide a self-explanation for the bolded sentence below:

Red blood cells have the vital role of carrying oxygen to all of the cells in the body. They also pick up waste carbon dioxide for removal. These cells are the most numerous of the blood cells. The disk shape of red blood cells results in a large surface area, which enables them to be efficient at gas diffusion.

Red blood cells contain a large, complex protein called hemoglobin. Hemoglobin binds to the oxygen and carbon dioxide that the red blood cells transport. Each red blood cell contains about 250 million hemoglobin molecules, each carrying four molecules of oxygen. Hemoglobin also contains iron, which gives blood its red color. Molecular oxygen can also be transported by another route, in dissolved blood plasma. However, oxygen is poorly soluble in water, so only about 1.5% is carried in dissolved form. Therefore, most oxygen is carried by hemoglobin.

Red blood cells lack a nucleus and the organelles found in other cells. **Therefore, these cells cannot reproduce or repair themselves.**

---

7. Please read the text and provide a self-explanation for the bolded sentence below:

Red blood cells have the vital role of carrying oxygen to all of the cells in the body. They also pick up waste carbon dioxide for removal. These cells are the most numerous of the blood cells. The disk shape of red blood cells results in a large surface area, which enables them to be efficient at gas diffusion.

Red blood cells contain a large, complex protein called hemoglobin. Hemoglobin binds to the oxygen and carbon dioxide that the red blood cells transport. Each red blood cell contains about 250 million hemoglobin molecules, each carrying four molecules of oxygen. Hemoglobin also contains iron, which gives blood its red color. Molecular oxygen can also be transported by another route, in dissolved blood plasma. However, oxygen is poorly soluble in water, so only about 1.5% is carried in dissolved form. Therefore, most oxygen is carried by hemoglobin.

Red blood cells lack a nucleus and the organelles found in other cells. Therefore, these cells cannot reproduce or repair themselves. Red blood cells live for about three or four months before being broken down in the spleen. Iron from the broken-down cells is returned to the bone marrow to be recycled into new hemoglobin.

**Sometimes blood does not transport enough oxygen, resulting in a condition called anemia.**

---

8. Please read the text and provide a self-explanation for the bolded sentence below:

Red blood cells have the vital role of carrying oxygen to all of the cells in the body. They also pick up waste carbon dioxide for removal. These cells are the most numerous of the blood cells. The disk shape of red blood cells results in a large surface area, which enables them to be efficient at gas diffusion.

Red blood cells contain a large, complex protein called hemoglobin. Hemoglobin binds to the oxygen and carbon dioxide that the red blood cells transport. Each red blood cell contains about 250 million hemoglobin molecules, each carrying four molecules of oxygen. Hemoglobin also contains iron, which gives blood its red color. Molecular oxygen can also be transported by another route, in dissolved blood plasma. However, oxygen is poorly soluble in water, so only about 1.5% is carried in dissolved form. Therefore, most oxygen is carried by hemoglobin.

Red blood cells lack a nucleus and the organelles found in other cells. Therefore, these cells cannot reproduce or repair themselves. Red blood cells live for about three or four months before being broken down in the spleen. Iron from the broken-down cells is returned to the bone marrow to be recycled into new hemoglobin.

Sometimes blood does not transport enough oxygen, resulting in a condition called anemia. **This makes a person feel tired and weak.**

9. Please read the text and provide a self-explanation for the bolded sentence below:

Red blood cells have the vital role of carrying oxygen to all of the cells in the body. They also pick up waste carbon dioxide for removal. These cells are the most numerous of the blood cells. The disk shape of red blood cells results in a large surface area, which enables them to be efficient at gas diffusion.

Red blood cells contain a large, complex protein called hemoglobin. Hemoglobin binds to the oxygen and carbon dioxide that the red blood cells transport. Each red blood cell contains about 250 million hemoglobin molecules, each carrying four molecules of oxygen. Hemoglobin also contains iron, which gives blood its red color. Molecular oxygen can also be transported by another route, in dissolved blood plasma. However, oxygen is poorly soluble in water, so only about 1.5% is carried in dissolved form. Therefore, most oxygen is carried by hemoglobin.

Red blood cells lack a nucleus and the organelles found in other cells. Therefore, these cells cannot reproduce or repair themselves. Red blood cells live for about three or four months before being broken down in the spleen. Iron from the broken-down cells is returned to the bone marrow to be recycled into new hemoglobin.

Sometimes blood does not transport enough oxygen, resulting in a condition called anemia. This makes a person feel tired and weak. Anemia can result from too little iron in the diet, loss of blood due to injury or menstruation, or various medical conditions. One type of anemia, called sickle-cell disease, is characterized by red blood cells that are sickle-shaped instead of disk-shaped. **The shape of the cells causes them to clog blood vessels, preventing oxygen from reaching muscles and other tissues.**

**Comprehension Questions**

Please answer the following questions about the text you just read.

1. How does sickle-cell disease get its name?

______________________________________________

2. Explain why blood plasma is a poor carrier of oxygen?

______________________________________________

3. Explain why the disk shape of red blood cells is advantageous for gas diffusion?

______________________________________________

4. What causes a person to feel weak and tired in anemia?

______________________________________________

5. How many oxygen molecules can be carried in each red blood cell?

______________________________________________

6. What are the critical elements of regular body cells that enable these cells to reproduce or repair themselves?

______________________________________________

7. In the production of hemoglobin, where does iron come from?

______________________________________________

8. How does sickle-cell disease cause anemia?

______________________________________________

# APPENDIX E

# GATES-MACGINITIE READING TEST

**Comprehension Section**

You will now complete a reading comprehension test. There will be several short reading passages, followed by questions. The first passage is an example. Please choose the best answer, then read the explanation that follows.

Sometimes - not very often - we get two full moons in one month. That second full moon is called a "blue moon." No one knows why. Now we say "once in a blue moon" to mean "once in a long time."

To be a "blue moon," the moon must be

o dark. (1)

o long. (2)

o blue. (3)

o full. (4)

The passage says that a second full moon is called a "blue moon." So to be a "blue moon," the moon must be full.

What is it that no one knows?

o What the name is. (1)

o Who uses the name. (2)

o Where the name came from. (3)

o What the name means. (4)

The passage says that no one knows why the second full moon is called a "blue moon." So no one knows where the name came from.

Now you will move on to the real questions. You have 12 minutes to complete the assessment, and you must spend at least 8 minutes before submitting. Press the right arrow to begin.

Any list of mutualistic relationships would be heavily weighted toward the highly organized, impersonal world of the insects. The story of ants protecting and "milking" their cattle-like aphids, for example, is well known. Much less common is evidence of mutualism among warm blooded vertebrates, and mutualistic relationships that cross taxonomic <u>class</u> lines, say between birds and mammals, are especially rare.

The passage mentions the relation between ants and aphids as an example of

o crossing taxonomic class lines. (1)

o insects being similar to people.  (2)

o an impersonal world.  (3)

o mutualism.  (4)

In this passage, class means a

o style.  (1)

o school group.  (2)

o social group.  (3)

o category.  (4)

The passage characterizes insect societies as

o ordered.  (1)

o highly motivated.  (2)

o small in scale.  (3)

o weighted.  (4)

A pulsar is thought to be a rapidly spinning neutron star. Such stars can arise from the gravitational collapse of a supernova's core. It is in conserving angular momentum as it shrinks to a diameter of only several kilometers that the neutron star attains its high rotational velocity. If the neutron star continuously emits a beam of electromagnetic radiation from a spot in the magnetized plasma overlying its surface, the beam is swept around like the beacon of a lighthouse. Such a radio beam, striking the earth with each revolution of the neutron star, can account for the observed radio-frequency pulsations.

A supernova's core becomes a neutron star because of

o rotation.  (1)

o gravity.  (2)

o pulsation.  (3)

o magnetized plasma.  (4)

A neutron star speeds up because it

o gets smaller.  (1)

o has a radio frequency.  (2)

o is magnetized.  (3)

o emits a beam.  (4)

Pulsars are thought to send out a radio beam from

o their magnetic poles.  (1)

o explosions in their interior.  (2)

o one place near their surface.  (3)

o the place where the beam strikes the earth.  (4)

What does like the beacon of a lighthouse describe?

o Radiation sent out by a pulsar.  (1)

o The star from which a pulsar is formed.  (2)

o Signals scientists send out to detect pulsars.  (3)

o The path of an object caught in a pulsar's gravity.  (4)

How often the beam from a pulsar strikes the earth depends on

o how far the pulsar is from the earth.  (1)

o how large the pulsar is.  (2)

o how fast the pulsar is spinning.  (3)

o how strong the pulsar's magnetic field is.  (4)

It is customary to place the date for the beginnings of modern medicine somewhere in the mid-1930s, with the entry of the sulfonamides and penicillin into the pharmacopoeia, and it is usual to ascribe to these events the force of a revolution in medical practice. This is what things seemed like at the time. Medicine was upheaved, revolutionized indeed. Therapy had been discovered for great numbers of patients whose illnesses had previously been untreatable. Cures were now available. As we saw it then, it seemed a totally new world. Doctors could now cure disease, and this was astonishing, most of all to the doctors themselves.

During the 1930s, what did people believe had happened in the field of medicine?

o A destructive trend.  (1)

o A dramatic change.  (2)

o A return to old practices.  (3)

o A slowing down.  (4)

Sulfonamides and penicillins made doctors feel

o confused.  (1)

o like scientists.  (2)

o old-fashioned.  (3)

o more confident.  (4)

In this passage, pharmacopoeia means

o a medical research laboratory.  (1)

o medical school textbooks.  (2)

o a school for pharmacists.  (3)

o a stock of available medicines.  (4)

According to the passage, who was most amazed by sulfonamides and penicillin?

o Sick patients.  (1)

o Doctors.  (2)

o Patients who had recovered.  (3)

o Pharmacists.  (4)

Stephen's mother and his brother and one of his cousins waited at the corner of quiet Foster Place while he and his father went up the steps and along the colonnade where the Highland sentry was parading. When they had passed into the great hall and stood at the counter Stephen drew forth his orders on the governor of the bank of Ireland for thirty and three pounds; and these sums, the moneys of his exhibition and essay prize, were paid over to him rapidly by the teller in notes and in coin respectively. He bestowed them in his pockets with feigned composure and suffered the friendly teller, to whom his father chatted, to take his hand across the broad counter and wish him a brilliant career in the after life.

The passage suggests that the building was

o hidden.  (1)

o crowded.  (2)

o impressive.  (3)

o hard to get into.  (4)

What had Stephen done?

o He had won a prize.  (1)

o He had carried out orders.  (2)

o He had sold a painting.  (3)

o He had had a brilliant career.  (4)

Why did the teller give the notes to Stephen rapidly?

o To get rid of Stephen.  (1)

o To show that he was not impressed.  (2)

o Because he was being efficient.  (3)

o Because Stephen's mother was waiting.  (4)

It was difficult for Stephen to

o act calmly.  (1)

o pass into the hall.  (2)

o give up the orders.  (3)

o leave his mother waiting.  (4)

The teller took Stephen's hand to

o greet him.  (1)

o congratulate him.  (2)

o give him confidence.  (3)

o show him how to handle money.  (4)

The Museum that Alexander the Great set up in Alexandria was in effect the first university in the world. As its name implies, it was dedicated to the service of the Muses. It was, however, a religious body only in form, in order to meet the legal difficulties of the endowment in a world that had never foreseen such a thing as a secular intellectual process. It was essentially a college of learned men engaged chiefly in research and record, but also to a certain extent in teaching.

Why was the Museum set up as a religious body?

o So money could be given to it.  (1)

o So people could come worship there.  (2)

o So priests could work there.  (3)

o So religion could be taught.  (4)

The museum was most like a

o temple.  (1)

o university.  (2)

o hospital.  (3)

o show.  (4)

Which answer best describes the Museum?

o Famed for its athletes.  (1)

o Ineffective.  (2)

o Pioneering.  (3)

o Entertaining.  (4)

All "symmetrical" organisms develop asymmetries. A fruit fly, no longer than the tip of a lead pencil, having developed while stuck to the inside of a glass culture vessel, has different numbers of sensory bristles on its left and right sides, some flies having more on the left, some more on the right. Moreover, this side-to-side variation is as large as the difference among different flies. But the genes on the left and right sides of a fly are the same, and it seems absurd to think that the temperature, humidity, or concentration of oxygen was different between left and right sides of the tiny developing insect. The variation between sides is a result of random events in the timing of division and movement of the individual cells that produce the bristles, so called-developmental noise.

Why does the author put symmetrical in quotation marks?

o It is a scientific term.  (1)

o It is a new word that the author made up.  (2)

o The author is referring to another author's use of the term.  (3)

o The usual meaning of the word is not completely accurate in this context.  (4)

In this passage, the vessel is

o a boat.  (1)

o a container.  (2)

o a vein or artery.  (3)

o a window.  (4)

The passage implies that differences such as that between right- and left-hand fingerprints could be a result of

o differences in genes.  (1)

o differences between individuals.  (2)

o symmetry.  (3)

o unpredictable variations in the way cells divide.  (4)

How does the number of bristles on the right side affect the number of bristles on the left.

o It has no effect.  (1)

o It makes the left side have fewer bristles.  (2)

o It makes the left side have an equal number of bristles.  (3)

o It makes the left side have more bristles.  (4)

Please select the word "variation".

o variation.  (1)

o fruit flies.  (2)

o concentration of genes.  (3)

o organism asymmetries.  (4)

Margaret had just gotten her first pair of sunglasses, perfect cat-eyes, and she was amazed at how much she could see. She lay in the scrub grass beneath a stand of cottonwoods, took them off, and watched the branches turn gauzy and familiar. Then she put the glasses back on, bracing a little for the barrage of detail. Thousands of leaves leaped out, trembling and hard-edged. The narrow river, a few yards away, turned crunchy-looking again. Bird sounds attached themselves to small shapes on high branches.

She didn't know when her vision had started to go seriously bad. It had been so gradual, this nearsightedness, that she hadn't noticed it for a while. At first, it seemed only that a luxurious vagueness had come into her life. Then it had begun to make her uneasy. But this sudden return of all the details was more than she really wanted. It was unnerving. It gave her the same feeling she got when someone explained how something scientific works - osmosis, say, or photosynthesis. The explanations crowded out her imagination and made her feel bleak with information.

What was Margaret not sure of?

o Why she had been feeling uneasy.  (1)

o When she started to need glasses.  (2)

o Whether her glasses were working properly.  (3)

o Why everything looked so different through glasses.  (4)

What had Margaret liked about not seeing well?

o She needed to imagine things.  (1)

o She didn't have to work.  (2)

o She could get people to explain things.  (3)

o She wasn't expected to understand science.  (4)

It seemed to Margaret that, when she wore the glasses, she had

o a feeling of luxury.  (1)

o a greater enjoyment of nature.  (2)

o too much information.  (3)

o a greater awareness of sounds.  (4)

The passage suggests that Margaret would have been happier with glasses that were

o weaker.  (1)

o smaller.  (2)

o like cat-eyes.  (3)

o more stylish.  (4)

In this passage, the word bracing means

o turning.  (1)

o pushing away.  (2)

o stimulating.  (3)

o getting ready.  (4)

I had planned to write chronologically, but then realized that, of course, I don't think chronologically. Writing a memoir is like fishing. You cast your line and you pull on it when a fish strikes, but you never know what will be on the other end, for the ocean is deep and is filled with marvelous creatures that do not break the surface in expected order. Nor do they swim under the waves with the whales leading and the minnows at the end of a long straight line. A memoir, like a fish, will not thrive under every discipline. Another way of putting this is that if you alphabetize the Iliad you will have approximately the Athens telephone book. When I think back, things don't line up, they stand out, so I will take them as they come, as once I took them as they came.

In this passage, the author explains why he

o decided to write about himself.  (1)

o waited so long to begin writing.  (2)

o included details that seem unimportant.  (3)

o changed his mind about how he would write.  (4)

What do the ocean creatures represent?

o Events in the author's life.  (1)

o People the author has known.  (2)

o All the words in the language.  (3)

o The dangers of looking into one's past.  (4)

In this passage, the word discipline means

o punishment.  (1)

o a field of study.  (2)

o rules by which something is organized.  (3)

o training that perfects mental or moral qualities.  (4)

The Athens telephone book is used as an example of something that is

o too long.  (1)

o impossible to read.  (2)

o orderly but boring.  (3)

o full of information.  (4)

When the author says "...as once I took them," he means that

o he was always eager to do things.  (1)

o he could stand up to any difficulty.  (2)

o he believed that he deserved what he got.  (3)

o he dealt with experiences as they happened.  (4)

Fresco involves painting into wet lime plaster with pigment mixed into limewater. The layer of calcium carbonate formed by the limewater binds the pigments to the plaster wall, and the mutual wetness of the pigment and the surface causes the color to dye the wall. This makes for a highly permanent decoration, as long-lived as the building itself. Permanence is the main advantage of fresco and is, of course, its own recommendation.

Michelangelo's *Creation of Adam*, like all other works on the ceiling of the Sistine Chapel in the Vatican, is an example of fresco painting. Since plaster cannot be rewet, once it is dry, the fresco artist never applies more plaster to his surface that he knows he can finish in a single day. Consequently, we can find places in this fresco where plaster joints occur. There is a seam where Adam's neck fits onto his body and another at the line between the torso and the legs. Adam is about twelve feet long, and it took Michelangelo three sessions to complete him.

In fresco painting, the pigment is first mixed into

o plaster.  (1)

o limewater.  (2)

o oil.  (3)

o the wet part of the wall.  (4)

Why are fresco paintings long lasting?

o The seams are strong.  (1)

o The pigment becomes part of the wall.  (2)

o The plaster is protected by the layer of pigment.  (3)

o The painting is protected by a layer of plaster.  (4)

About how long does the plaster stay wet enough to paint?

o Ten minutes.  (1)

o An hour.  (2)

o A day.  (3)

o A week.  (4)

A fresco artist must be careful to

o rewet the plaster as needed.  (1)

o apply the plaster to small enough areas.  (2)

o let the plaster dry before beginning to paint.  (3)

o let the paint dry before applying plaster.  (4)

A seam in a fresco is a line

o where the wall joins the ceiling.  (1)

o between different colors.  (2)

o between areas painted at different times.  (3)

o where material has been added to strengthen the plaster.  (4)

The example of *Creation of Adam* shows how one can tell

o where the artist applied plaster.  (1)

o how long ago the fresco was painted.  (2)

o how large the figures on a ceiling fresco are.  (3)

o how many sessions it took to do a fresco.  (4)

In later life, John Quincy Adams recalled an incident typical of his mother Abigail's bravery and resourcefulness. In 1775 British troops from Boston were advancing on Braintree, searching for rebel arsenals. All day neighbors traveled the road in front of the Adams' farmhouse, retreating from the expected attack. Abigail was alone in her home with her children. When rebel troops arrived, they advised Abigail to flee. Instead she stayed, handing over all her precious pewter to the rebels, helping them melt down the metal for bullets. The rebel soldiers departed, and Abigail remained, expecting the worst but refusing to give in to the panic that possessed some of her neighbors. "Do you wonder," wrote her son, "that a boy of seven who witnessed this scene is a patriot?"

The neighbors who passed the Adams' house were trying to

o defend their homes.  (1)

o avoid being hurt.  (2)

o join one of the armies.  (3)

o get to Boston.  (4)

The passage suggests that the rebels had little o ammunition.  (1)

o concern for Abigail.  (2)

o knowledge of the countryside.  (3)

o warning that the British were advancing.  (4)

What demonstrated Abigail's resourcefulness was the way she

o fooled the British troops.  (1)

o sent messages to the rebel troops.  (2)

o learned where the British troops had come from.  (3)

o provided what was needed from what she had available.  (4)

John Quincy Adams believed that this experience was a source of his

o resourcefulness.  (1)

o interest in military history.  (2)

o courage.  (3)

o love of country.  (4)

A crowd of people surged in to the Eighth Avenue express at 59th Street. By elbowing other passengers in the back, by pushing and heaving, they forced their bodies into the coaches, making room for themselves where no room had existed before. As the train gathered speed for the long run to 125th Street, the passengers settled down into small private worlds, thus creating the illusion of space between them and their fellow passengers. The worlds were built up behind newspapers and magazines, behind closed eyes or while staring at the varicolored show cards that bordered the coaches.

Why was it difficult to get on the train?

o The train didn't stop long enough.  (1)

o There was a barrier in the way.  (2)

o The train was already full.  (3)

o The people were reading newspapers.  (4)

The newspapers helped the passengers

o pass the time more quickly.  (1)

o forget where they were going.  (2)

o sleep.  (3)

o feel that they were by themselves.  (4)

Staring at the show cards served the same purpose as

o finding a seat.  (1)

o getting on the train.  (2)

o shutting one's eyes.  (3)

o staring at other passengers.  (4)

In this passage, the word <u>run</u> means

o trip.  (1)

o race.  (2)

o string of good luck.  (3)

o series of performances.  (4)

**Vocabulary Section**

You will now complete a vocabulary assessment. Each question will have a sentence or phrase that contains an underlined word. Please select the answer choice that is closest in meaning to the underlined word. The first two questions are examples. Please select the correct answer, then read the explanation that follows.

a big <u>garage</u>

o place for cars  (1)

o machine  (2)

o sidewalk  (3)

o covered porch  (4)

o cloth sack  (5)

The answer that means most nearly the same as **garage** is **place for cars** - a big **place for cars** means the same as **a big garage**. So the answer is **place for cars**.

They will <u>close</u> it.

o stay near  (1)

o begin (2)

o make (3)

o shut (4)

o go past (5)

**Shut** means most nearly the same as **close**. So you should have selected the answer **shut**.

Now you will move on to the real questions. You have 7 minutes to complete the assessment. Press the right arrow to begin.

They can recline.

o lie back (1)

o turn over (2)

o grow weaker (3)

o rebel (4)

o look around (5)

He must crave it.

o cut (1)

o try to avoid (2)

o crank (3)

o want (4)

o reach for (5)

a pesky animal

o rare (1)

o stray (2)

o disease-carrying (3)

o lucky (4)

o troublesome (5)

a different stanza

o tune (1)

o verse (2)

o volume (3)

o row of seats (4)

o stamp (5)

the peculiar person

o careful (1)

o worried (2)

o dangerous (3)

o strange (4)

o impeccable (5)

the expected climax

o correct answer (1)

o weather (2)

o high point (3)

o mountain climber (4)

o parallax (5)

a pretty damsel

o dressing gown (1)

o castle (2)

o dance (3)

o young woman (4)

o flower (5)

the better lounge

o couch (1)

o leash (2)

o blast-off (3)

o blanket (4)

o soft floor covering (5)

a good basis

o guess (1)

o experience (2)

o foundation (3)

o audience (4)

o emphasis (5)

They are somber.

o soggy (1)

o asleep (2)

o thick (3)

o uneven (4)

o gloomy (5)

a big hindrance

o rear axle (1)

o obstacle (2)

o loss of money (3)

o contest (4)

o back entrance (5)

She will be ostracized.

o excluded from the group (1)

o arrested (2)

o put in a trance (3)

o worn out (4)

o yelled at  (5)

the clear grounds

o gains  (1)

o reasons  (2)

o sayings  (3)

o decisions  (4)

o moans  (5)

They were dilated.

o cleaned  (1)

o pleased  (2)

o made larger  (3)

o paid off  (4)

o diverted  (5)

a successful debut

o opera singer  (1)

o collection of a debt  (2)

o concert  (3)

o argument  (4)

o first appearance  (5)

a small gully

o cloth sack  (1)

o artificial lake  (2)

o wagon  (3)

o trench  (4)

o clay jug  (5)

He did it readily.

o carelessly (1)

o easily (2)

o realistically (3)

o by the book (4)

o after all (5)

It should be amended.

o explained (1)

o praised (2)

o asked for (3)

o returned (4)

o corrected (5)

They may be erected.

o voted in (1)

o built (2)

o erased (3)

o sent away (4)

o painted (5)

It might befit them.

o behold (1)

o punish (2)

o surprise (3)

o suit (4)

o show (5)

They can evade it.

o invade (1)

o grade (2)

o chase after  (3)

o cover  (4)

o escape  (5)

the welcome legacy

o inheritance  (1)

o promotion  (2)

o law  (3)

o job  (4)

o prize  (5)

Please select the word "apple".

o jam  (1)

o apple  (2)

o caramel  (3)

o tortilla  (4)

o pear  (5)

They want to woo him.

o express concern for  (1)

o inspire  (2)

o try to persuade  (3)

o scare  (4)

o save  (5)

a strong fulcrum

o balance beam  (1)

o hawk  (2)

o vacuum  (3)

o crane  (4)

o support for a lever (5)

It was inert.

o filled with gas (1)

o dry (2)

o not active (3)

o without purpose (4)

o not wanted (5)

the last hurdle

o faint hope (1)

o difficulty (2)

o fall (3)

o group (4)

o short race (5)

She was debilitated.

o weakened (1)

o in debt (2)

o irritated (3)

o lied to (4)

o embarrassed (5)

a great torrent

o rushing stream (1)

o tortoise (2)

o snowfall (3)

o valley (4)

o vacant space (5)

the old trowel

o soft cloth (1)

o trick (2)

o garden tool (3)

o steep path (4)

o fairy-tale creature (5)

the shiny vial

o flower vase (1)

o visor (2)

o faucet (3)

o little bottle (4)

o stone (5)

They may contend.

o continue (1)

o attend (2)

o appear (3)

o compete (4)

o join together (5)

the small urchin

o scamp (1)

o bully (2)

o sad face (3)

o chain (4)

o city house (5)

a large receptacle

o closet (1)

o pair of glasses (2)

o container  (3)

o amount of credit  (4)

o dump  (5)

the real anguish

o love  (1)

o anger  (2)

o difficulty  (3)

o slant  (4)

o sorrow  (5)

He did it lackadaisically.

o cheerfully  (1)

o without help  (2)

o without understanding  (3)

o incorrectly  (4)

o without interest  (5)

the old beacon

o flagpole  (1)

o signal light  (2)

o watchman  (3)

o food  (4)

o airport  (5)

They may straggle.

o choke  (1)

o fight  (2)

o come back  (3)

o stray  (4)

o tingle (5)

He will wield authority.

o hold and use (1)

o break with (2)

o take back (3)

o depend on (4)

o hide from (5)

They are deluged.

o swamped (1)

o scared away (2)

o turned over (3)

o put in sacks (4)

o delighted (5)

the last jaunt

o swim (1)

o bus (2)

o expense (3)

o leisure trip (4)

o connecting part (5)

They will expunge it.

o say (1)

o not see (2)

o be ready for (3)

o arrange (4)

o erase (5)

the long veranda

o porch  (1)

o hallway  (2)

o sofa  (3)

o path  (4)

o verse of a poem  (5)

an annoying nonchalance

o silence  (1)

o facial expression  (2)

o lack of concern  (3)

o personality  (4)

o way of talking  (5)

It had been defiled.

o trimmed off  (1)

o made definite  (2)

o recorded  (3)

o explained  (4)

o made dirty  (5)

an old statute

o work of art  (1)

o law  (2)

o boundary  (3)

o building  (4)

o toll road  (5)

# APPENDIX F

## ENJOYMENT AND FEEDBACK SURVEY

Please view the following statements about the game you just played and respond with how strongly you agree or disagree.

| | Strongly disagree | Disagree | Neither agree nor disagree | Agree | Strongly agree |
|---|---|---|---|---|---|
| This was fun to complete. | o | o | o | o | o |
| This was frustrating. | o | o | o | o | o |
| I enjoyed completing the tasks in this. | o | o | o | o | o |
| This was boring. | o | o | o | o | o |
| This was easy. | o | o | o | o | o |
| I would do this again. | o | o | o | o | o |

Please respond with any specific feedback you have about the game, so that we can improve it for future use.

____________________________________________________________

# APPENDIX G

# USER EXPERIENCE SURVEY

Please provide any additional feedback regarding your experiences when playing the games. We are looking for honest feedback to help with improving the games.

________________________________________________________________

For all the following items, please score how you feel right now, just after interacting with the games.

| | Not at all (1) | Slightly (2) | Moderately (3) | Very (4) |
|---|---|---|---|---|
| Relieved (1) | o | o | o | o |
| At Ease (2) | o | o | o | o |
| Nervous (3) | o | o | o | o |
| Satisfied (4) | o | o | o | o |
| Fed up (5) | o | o | o | o |
| Fine (6) | o | o | o | o |
| Worried (7) | o | o | o | o |
| Confident (8) | o | o | o | o |
| Annoyed (9) | o | o | o | o |

| | | | | |
|---|---|---|---|---|
| Concerned (10) | o | o | o | o |

1. How much did you enjoy the games today?
    1. Not at all
    2. Somewhat
    3. A lot
    4. A great deal
2. How much attention did you devote to interacting with the games in this study?
    1. Not much attention
    2. Some attention
    3. Much attention
    4. Very much attention
3. Please select the fourth option.
    1. Not at all.
    2. Somewhat.
    3. A lot.
    4. A great deal.
4. How difficult did you find interacting with the games in this study?
    1. Not at all difficult
    2. Not so difficult
    3. Difficult
    4. Very difficult
5. How useful do you consider these games?
    1. Not at all useful
    2. Not so useful
    3. Useful

4. Very useful

# APPENDIX H

## MAP CONQUEST DISTRIBUTIONS OF NLP INDICES

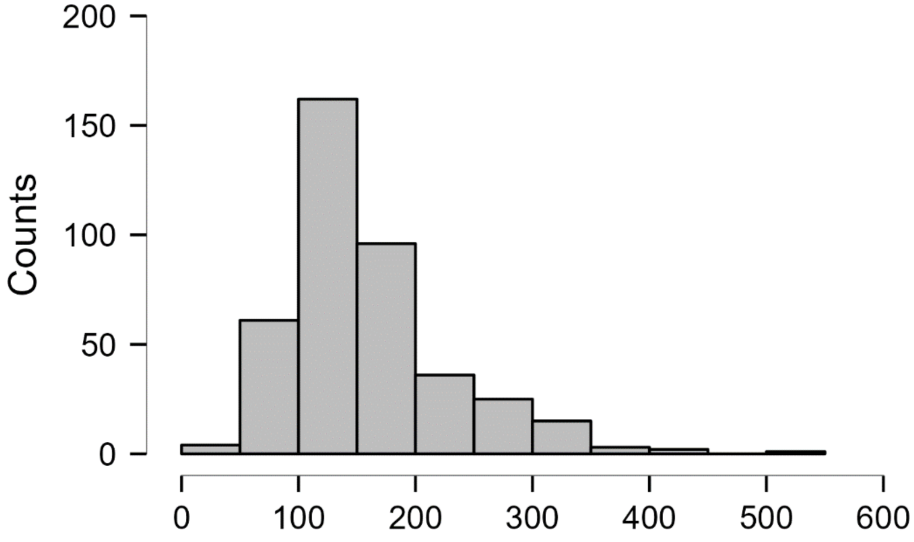


*Figure H1*. Frequency distribution of number of words used overall within a participant's self-explanations.

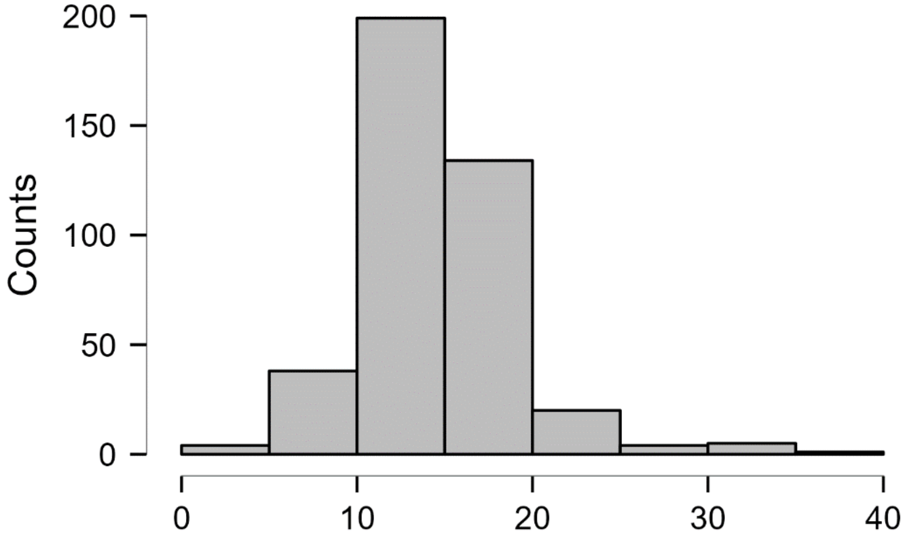


*Figure H2*. Frequency distribution of average sentence length, in number of words.

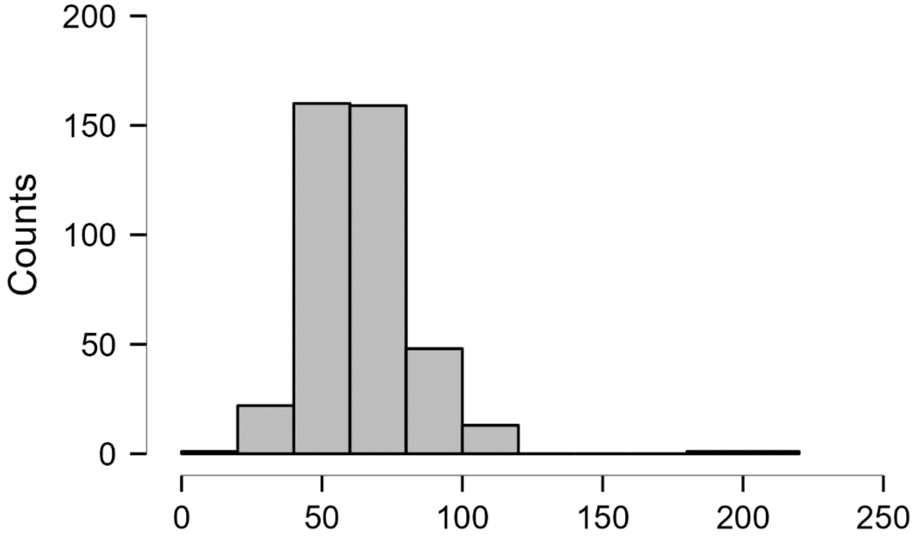


*Figure H3*. Frequency distribution of lexical diversity.

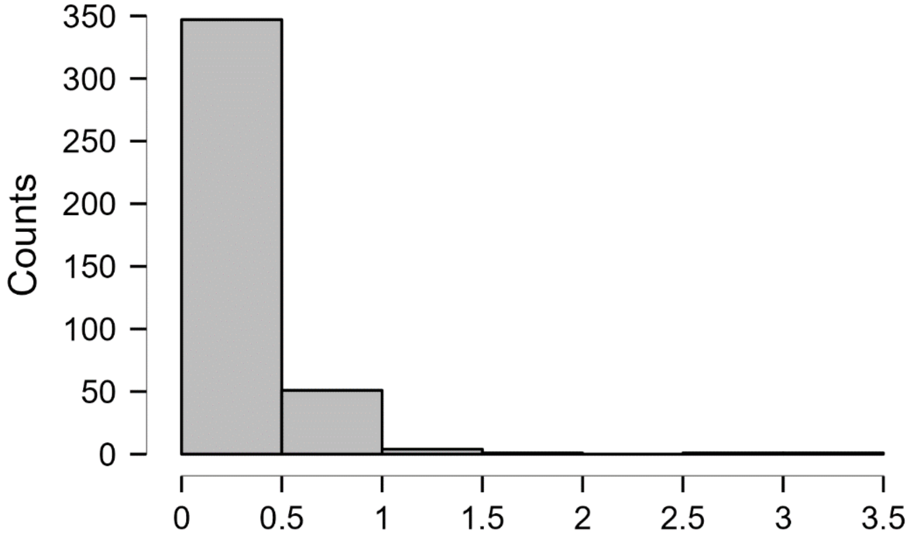


*Figure H4*. Frequency distribution of causal ratio.

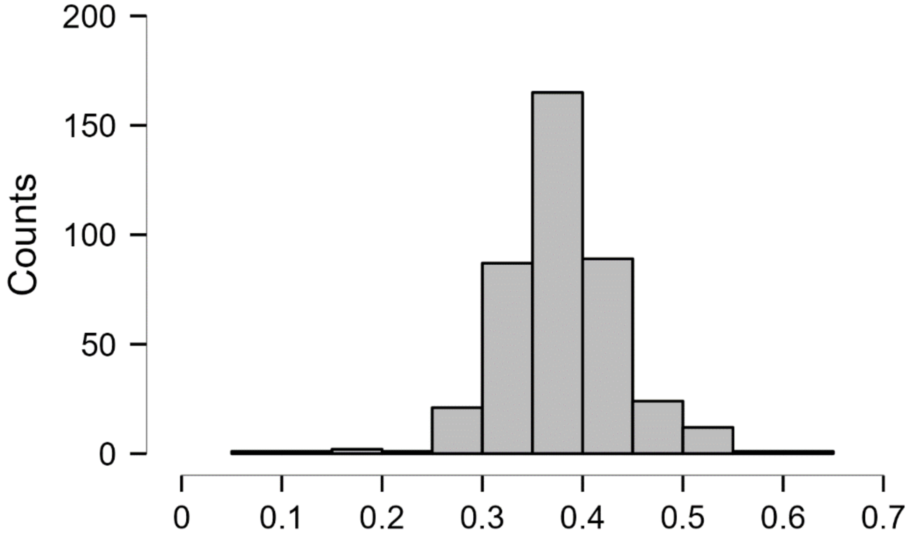


*Figure H5*. Frequency distribution of givenness.

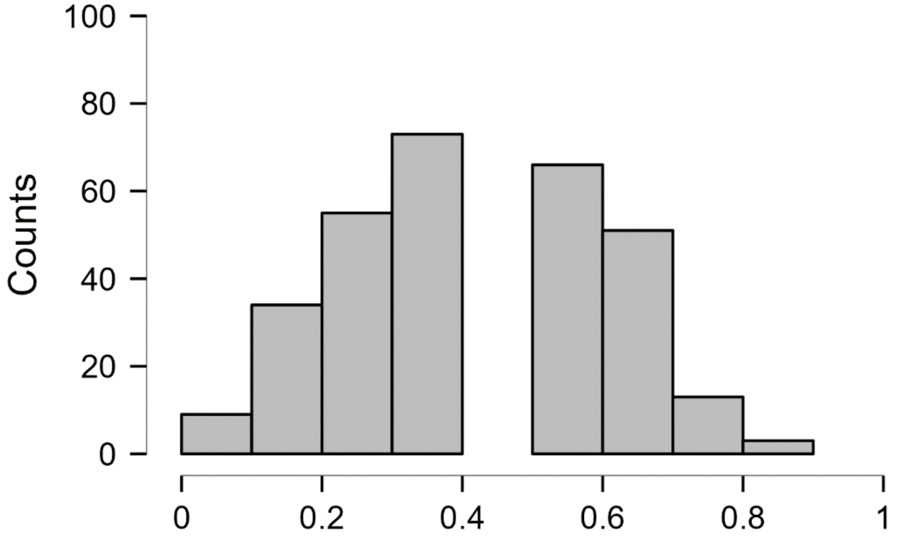


*Figure H6*. Frequency distribution of verb overlap between self-explanations.

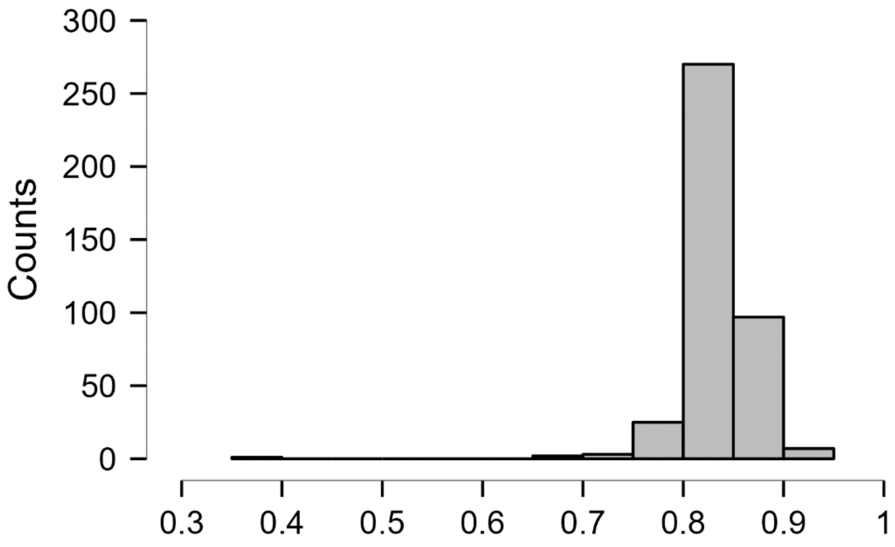


*Figure H7*. Frequency distribution of semantic overlap between self-explanations

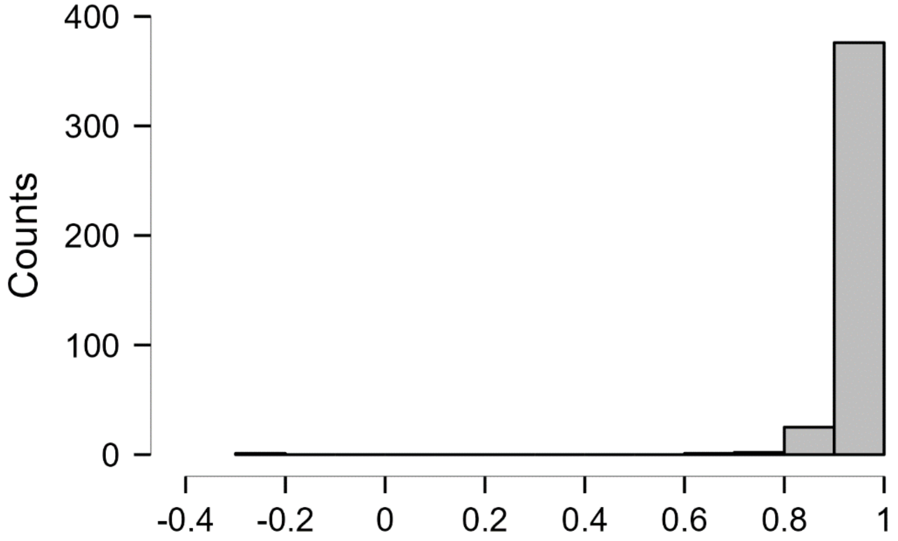


*Figure H8*. Frequency distribution of semantic overlap between the participants' self-explanations and the source text.

# APPENDIX I

## CORRELATIONS OF GAME ENJOYMENT AND PERFORMANCE

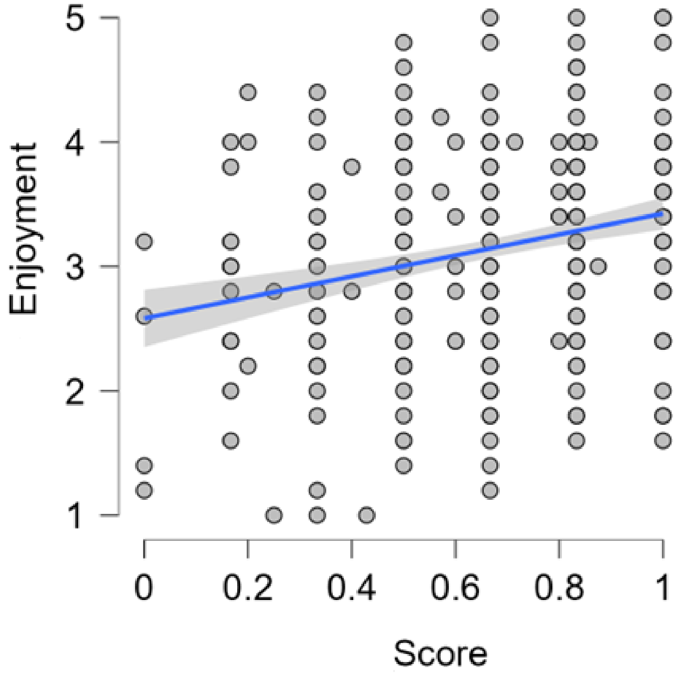


*Figure I1*. Correlation of Paraphrase Quest score with Paraphrase Quest enjoyment (r = .25).

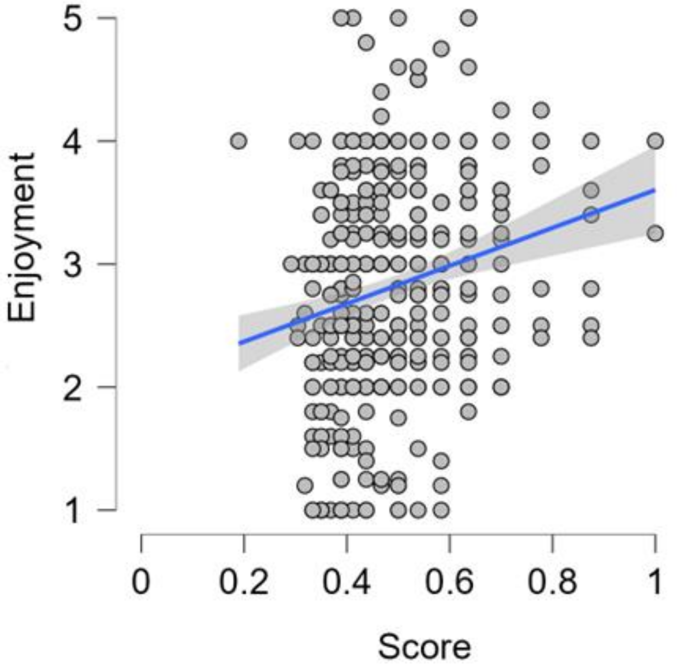


*Figure I2*. Correlation of Fix It score with Fix It enjoyment (r = .21).

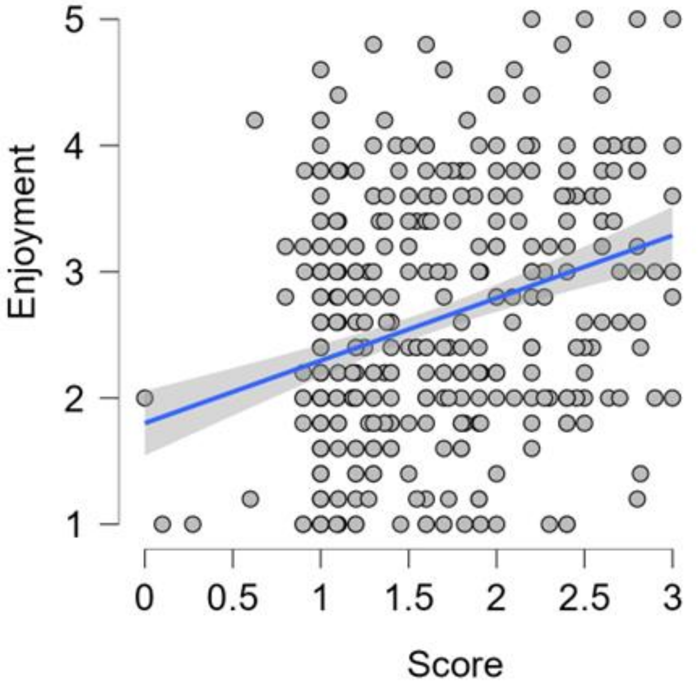


*Figure I3*. Correlation of Map Conquest average system score across self-explanations with Map Conquest enjoyment (r = .31).

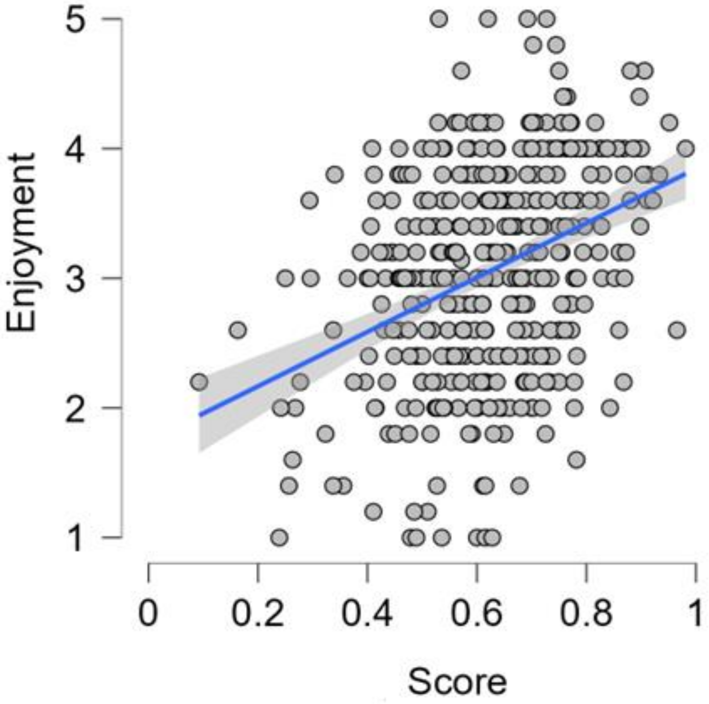


*Figure I4*. Correlation of Vocab Flash score with Vocab Flash enjoyment (r = .37).

# APPENDIX J

# IRB APPROVAL

ASU Knowledge Enterprise Development

APPROVAL: MODIFICATION

Danielle McNamara
EVPP: Executive Vice President and Provost, Office of the
480/727-5690
Danielle.McNamara@asu.edu

Dear Danielle McNamara:

On 10/28/2022 the ASU IRB reviewed the following protocol:

| | |
|---|---|
| Type of Review: | Modification / Update |
| Title: | Game-Based Literacy Assessment |
| Investigator: | Danielle McNamara |
| IRB ID: | STUDY00011488 |
| Funding: | Name: DOD: Navy |
| Grant Title: | None |
| Grant ID: | None |
| Documents Reviewed: | • Data request communication UTO and Provost, Category: Other;<br>• ONR_GBLA_Protocol_9-7-22.docx, Category: IRB Protocol;<br>• ONR_SONA_Consent_9-7-22.pdf, Category: Consent Form;<br>• ONR_SONA_Posting_9-2-22.pdf, Category: Recruitment Materials;<br>• ONRSampleQuestions_9-2-22.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); |

The IRB approved the modification.

When consent is appropriate, you must use final, watermarked versions available under the “Documents” tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).
Sincerely,

IRB Administrator

cc: Tracy Arner
Srikanth Ganji
Ruihao Zhang
Jason Miller
Reese Butterfuss
Ying Fang
Karen Taylor
Tracy Arner
Revanth Chimmani
Michelle Banawan
Danielle McNamara
Cecile Perret
Emily Goblirsch
Jianmin Dai
Renu Balyan
Tong Li
Rod Roscoe
Dian Apu
Katerina Christhilf
Linh Huynh
Micah Watanabe
Natalie Newton