2019

The Impact of Task Difficulty on Reading Comprehension Intervention with Computer Agents

Ying Fang

THE IMPACT OF TASK DIFFICULTY ON READING COMPREHENSION INTERVENTION WITH COMPUTER AGENTS

By

Ying Fang

A Dissertation

Submitted in Partial Fulfillment of the

Requirement for the Degree of

Doctor of Philosophy

Major: Psychology

The University of Memphis

December 2019

Acknowledgements

My deepest and sincere appreciation first goes to my two advisors Professor Xiangen Hu and Professor Arthur C. Graesser. I want to thank them for their consistent support, motivation, encouragement, and wisdom. Without their guidance, this dissertation would not have been possible.

I would like to extend my gratitude to my committee members, Professor Philip Pavlik and Professor Su Chen for their valuable suggestions, and ideas which gave me different perspectives and broadened my research horizon.

My thanks also go to Zhiqiang Cai, a brilliant mathematician whom I learned so much from, and was always willing to help. Without his support, it would not have been possible to conduct this dissertation research.

I would also like to thank my coworkers Keith Shubeck, Anne Lippert and Colin Carmon for their assistance and insightful discussions.

Most importantly, I would like to thank my husband, Wehcaho Zhang and my daughter, Cindy Zhang for their support during these years. Without their understanding and encouragement, I could not imagine how this dissertation would have been possible.

The research was supported by the Institute of Education Sciences, US Department of Education, through Grant R305C120001. Any opinions, findings, and conclusions, or recommendations expressed in this dissertation are those of the author and do not represent views of the Institute or the US Department of Education.

Abstract

Fang, Ying. Ph.D. The University of Memphis. September 2019. The Impact of Task Difficulty on Reading Comprehension Intervention with Computer Agents. Major Professor: Xiangen Hu, Ph.D.

According to Vygotsky's zone of proximal development (ZPD), students benefit from tasks that are difficult but can be achieved with the guidance of a skilled partner. This concept has often been implemented during the development of intelligent tutoring systems (ITSs), but there remains a need to empirically investigate the validity of this concept in reading comprehension interventions. Individual differences need to be considered in order to assign tasks to individual students with the right level of difficulty. Mixed results have emerged from previous research that specifically investigated the potential interaction between student attributes and ITSs on learning in the domain of reading comprehension. The present dissertation explored how and to what extent individual differences interact with task difficulty when assessing performance during learning, memory, and engagement during a reading comprehension intervention with an ITS. A within-subject experiment was conducted which involved two different conditions: (1) an increasing difficulty order condition in which easy learning tasks were followed by difficult tasks versus (2) a decreasing difficulty order condition in which difficult learning tasks were presented before easy tasks. The students' learning performance (i.e., accuracy and time to answer questions) was tracked during their interactions with AutoTutor, a conversation-based ITS. The students' reading skills were assessed by a reading comprehension test whereas memory was assessed by a recognition test on content delivered by AutoTutor. Results indicated that students with lower reading skills had better memory for surface structure (i.e., wording and syntax) in the increasing difficulty order condition, whereas students with higher reading skills had better memory for surface structure in the decreasing difficulty order condition. Learning performance during AutoTutor was not significantly different between the two task difficulty order conditions, but engagement as indicated by reading time was found to affect high-skill and low-skill students' memory differently in the increasing and decreasing task difficulty order conditions. Implications for improving the adaptivity of ITSs are discussed.

Table of Contents

List of Tables

List of Figures

Chapter 1: Goals and Scope of the Dissertation

Context of Problem

Task difficulty of an instructional activity can be defined as the degree to which the activity requires a considerable amount of cognitive or physical effort in order to develop a student's knowledge or skills (Orvis, Horn, & Belanich, 2008). Learning tasks should not be too hard or too easy, but at the right level of difficulty given a student's skill level or prior knowledge, according to the Goldilocks Principle (Graesser, 2009; Landauer & Psotka, 2000). The Goldilocks Principle is in line with Vygotsky's zone of proximal development (ZPD), according to which the tasks should be too difficult for a student to master on his/her own but can be achieved with the guidance and encouragement from a skilled partner (Vygotsky, 1978). Individuals are challenged when they are given a task that demands skills or knowledge beyond their current capabilities (Van Velsor & McCauley, 2004). Individuals are allegedly motivated more by challenging tasks that provide an intermediate probability of success, than they are with tasks with a higher certainty of success or failure (Belanich, Sibley, & Orvis, 2004; Malone & Lepper, 1987).

The concept of ZPD has been implemented during the development of intelligent tutoring systems (ITSs). One question researchers ask when examining the effectiveness of ITSs is how individual differences interact with learning in ITSs, an approach rooted in the search for aptitude-treatment interactions (ATI) (Cronbach, 1957; Snow, 1991). Interestingly, no consensus has been reached and mixed results have been reported on ATI's for different ITSs and domains. For example, research on ITSs efficacy that has examined students' prior knowledge have reported contradictory results including: low knowledge students benefit most from ITSs (Beal, Arroyo, Cohen, Woolf & Beal, 2010; Beal, Walles, Arroyo, & Woolf, 2007), intermediate knowledge students benefit more than low knowledge students (Steenbergen-Hu & Cooper, 2013), and high knowledge students benefit more than medium and low knowledge students (Ma Adesope, Nesbit, & Liu, 2014).

In the domain of reading comprehension, previous research also indicated inconsistent and even conflicting results on how a variety of individual differences interact with ITSs on learning (Ji et al., 2018; Meyer & Lei, 2012; Meyer et al., 2010; Wijekumar, Meyer, & Lei, 2017). For instance, Meyer and Lei (2017) reported that female students benefited more from an ITS than male students, but Wijekumar and her colleagues (2017) did not find a gender difference among students using the same ITS. Ji et al. (2018) found lower skilled readers benefited the most from an ITS, whereas Wijekumar, Myer and Lei (2012) did not observe any difference between low, medium, and high skilled readers using the same ITS.

In addition to individual differences, another factor found to affect reading comprehension is text cohesion. Text cohesion was found to interact with individual differences for reading comprehension. For example, readers with low prior knowledge were found to learn better with high-cohesion texts, whereas readers with high prior knowledge benefited more from low-cohesion texts (McNamara, Kintsch, Songer, & Kintsch, 1996; O'reilly & McNamara, 2007; O'Reilly, Sinclair, & McNamara, 2004). Text cohesion was also found to interact with reading skill. Voss and Silfies (1996) reported that the effect of text cohesion on comprehension depended on reading skill. Their correlational analyses indicated that reading skill was primarily correlated with better performance on high-cohesion texts, whereas prior knowledge correlated better with improved performance on low-coherent texts. Ozuru, Dempsey, & McNamara (2009) reported a three-way interaction between prior knowledge, reading skill, and text cohesion which indicated the degree to which students benefited from more coherent texts dependent on the readers' reading skill. Specifically, less-skilled, high-knowledge readers benefited from low-cohesion texts; skilled high-knowledge readers gained more from more coherent texts. In summary, the research that has looked at ATIs in ITSs for reading comprehension is sparse and inconsistent.

Purpose of the Study

This dissertation draws its motivation from the Zone of Proximal Development principle, the hunt for ATI's, and empirical studies on the efficacy of ITSs. This dissertation investigated the effect of task difficulty order on memory, engagement and performance during learning in a reading comprehension intervention with an ITS, while also considering individual differences. Task difficulty was based on the difficulty of the texts that students read. Text difficulty was scaled by an automated text analysis tool called Coh-Metrix (Graesser, McNamara, & Kulikowich, 2011; Graesser et al., 2014). There were two task difficulty order conditions: an increasing difficulty condition and a decreasing difficulty condition. In the increasing difficulty order condition, easy texts were followed by difficult tasks; in the decreasing difficulty order condition, difficult texts were presented before easy tasks. The learning materials and tasks were provided by AutoTutor, a conversation-based ITS (Graesser et al., 2004). Students were guided through learning tasks which involved reading texts and answering questions asked by computer agents periodically. AutoTutor records the students' answers to questions and their response time, which were used as measures of performance. The students' reading skills were measured by comprehension scores and reading fluency, as assessed by a reading comprehension test that has been validated psychometrically, i.e., the maze subtest (Sabatini, Bruce, Steinberg, & Weeks, 2019; Wayman, Wallace, Wiley, Ticha, & Espin, 2007). The students' memory was assessed by a recognition test on content delivered by AutoTutor. Engagement during learning was inferred from text reading time and question response time recorded by AutoTutor. Individual differences in reading skill were explored in order to examine the extent to which they affected memory, performance during learning, and engagement along with task difficulty order.

Specific Research Questions

This dissertation investigated the following questions: (1) Does task difficulty order (i.e., increasing difficulty versus decreasing difficulty) impact memory, engagement and performance during learning in a reading comprehension intervention with the AutoTutor ITS? (2) Do individual differences interact with task difficulty order in terms of their impact on memory, engagement, and performance during learning? If yes, how and to what extent do individual differences interact with task difficulty order? (3) Does engagement moderate the effect of task difficulty order on memory and performance during learning?

Significance of the Study

This dissertation contributes to the learning sciences in three ways. First, the current study investigated the impact of task difficulty order, which has not been fully explored in the domain of reading comprehension. The effect of task difficulty on memory or performance during learning has been studied in various ways (McNamara et al., 1996; McDaniel, Einstein, Dunay, & Cobb, 1986; O'Brien & Myers, 1985; Weaver & Bryant, 1995). However, it is uncommon to manipulate the order of task difficulty in a controlled study, particularly within the reading comprehension domain. This gap may be partially attributed to the challenge of measuring task difficulty automatically and objectively. In this study, a linguistic computational tool called Coh-Metrix was used to analyze task difficulty so that the task difficulty order can be manipulated. Specifically, task difficulty was scaled on a text difficulty metric computed by Coh-Metrix (Graesser & McNamara, 2011; Graesser et al., 2011; McNamara, Graesser, McCarthy, & Cai, 2014). Graesser et al. (2014) identified a composite measure called formality, based on the component measures produced by Coh-Metrix. This formality measure from Coh-Metrix was used to determine the difficulty of each text and its associated question answering items.

A second contribution of the dissertation is that the tasks were situated in an ITS with two conversational agents. The computer agents (i.e., a teacher agent and a peer agent) guided students through their process by having conversations with them. Conversations in AutoTutor have yielded significant improvements within a variety of domains, including physics, computer literacy, critical thinking (Graesser, 2016; Nye, Graesser, & Hu, 2014). However, it is not clear how and to what extent students benefit form trialogues (i.e., three-way conversation with a tutor agent and peer agent) in the domain of reading comprehension. The findings in this dissertation will prove useful for future ITSs design, particularly regarding how the system should order reading tasks to adapt to individual students.

The third contribution is the exploration of aptitude-treatment interactions (ATI) in this study. Specifically, this dissertation explored how individual differences interacted with task difficulty order in terms of their effect on memory, engagement and performance during learning. Individual differences have been found to interact with the effects of learning environments in the domain of math and science (Beal et al., 2010; Craig et al., 2013; Ma et al., 2014; Steenbergen-Hu & Cooper, 2013). In the domain of reading comprehension, only a few studies considered individual differences when evaluating the effectiveness of educational software or ITSs. The results of the few studies that did consider individual differences are mixed, so it remains unclear as to how individual differences interact with ITSs (Ji et al., 2018; Meyer et al., 2010; Wijekumar et al., 2012; Wijekumar et al., 2017). This dissertation took individual differences into account and investigated the interaction between individual differences and task difficulty order, a relatively underexplored theme in ITSs research. If the interaction between reading skill and task difficulty order is found to impact students' memory, engagement, and performance during learning, there is some useful guidance on how to improve the adaptivity of ITSs.

Chapter 2 of this dissertation introduces Zone of Proximal Development (ZPD) and Aptitude-Treatment Interactions (ATI) and their applications in the design and research of ITSs. Chapter 2 also reviews ITSs studies in the domain of reading comprehension, ending with an elaboration of the hypotheses of the present study. Chapter 3 describes the experimental design and methods, including participants, design, materials, measures, and procedures. Chapter 4 reports the results of the analyses on the three research questions. Finally, Chapter 5 summarizes the findings and discusses the implications and limitations of the dissertation.

Chapter 2 Introduction

Zone of Proximal Development (ZPD)

The term "Zone of Proximal Development (ZPD)" was first introduced by Soviet psychologist Lev Vygotsky in his book *Mind in Society* (1978). Vygotsky described ZPD as, "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978, p. 86). Vygotsky's ZPD expanded on Jean Piaget's theory that children are autonomous learners (Schaffer, 2006). Although Vygotsky agreed that natural and spontaneous knowledge development was important, he also believed that it was crucial for children to interact with more knowledgeable others. Vygotsky argued against the use of knowledge-based tests to determine a student's intelligence. Instead, he suggested that intelligence is better measured by observing a student's ability to solve problems independently and their ability to solve problems with assistance from an adult (Berk & Winsler, 1995). The definition of ZPD was eventually expanded to include learning from peers with higher skill sets (Penguin *Dictionary of Psychology*, 2009).

Historically, ZPD has been applied in two major ways, namely dynamic assessment and interactive learning that emphasizes scaffolding (Winegar, 1988). Regarding dynamic assessment, Wenegar suggested ZPD should be interpreted as a child's readiness, a knowledge state that changes over time. For example, when a child is given a task along with hints for how to solve it, the child moves towards the solution. Afterwards, the child can be observed to transfer the strategy to new tasks. The efficiency of this transfer from one task to another is an indicator of the student's ZPD. Dynamic assessment seeks to provide accurate information about an individual's current learning ability and learning processes which can be used to develop individualized teaching strategies. Dynamic assessments are contrasted by static assessments (e.g., IQ tests), which assume fixed and immutable characteristics of intelligence or cognitive functioning (Feurerstein, Falik, Rand, & Feurerstein, 2002).

ZPD is also found in interactive learning that emphasizes scaffolding. Although Vygotsky never specifically mentioned “scaffolding”, this term was introduced when researchers applied his concept of ZPD to educational contexts (Schaffer, 2006). Balaban (1995) defined scaffolding as a way the adult guides the child’s learning through focused questions and positive interactions. Winegar (1988) explained the timing for scaffolding with an example. He suggested that if a given task can be accomplished by the sequence of X-Y-Z, and the child’s ability of performing Y is not mature yet, the tutor needs to scaffold the child to accomplish Y and make it possible for the child to solve the task. When the child’s ability of solving Y is mature, the child can finish the task independently and the scaffolding of the tutor is no longer needed. More recent research investigating tutor actions suggested scaffolding steps should include: orienting the student by describing the problem, providing the student a goal, and helping the student complete a reasoning step (Chi, 1996; Graesser, Person, & Magliano, 1995). Chi and Wylie (2014) explained scaffolding is effective because it requires students to be generative and respond constructively. For instance, the scaffolding questions like “what else” encourages students to generate an active response rather than yes/no response.

Intelligent Tutoring Systems and the Application of ZPD

Intelligent tutoring systems (ITSs) are a specific type of electronic learning environments that intelligently adapt to individual students, implement pedagogical strategies, and sometimes resemble human tutoring (Graesser, Conley, & Olney, 2012). The architecture of ITSs has four major components: the domain model, the student model, the pedagogical model and the user interface (Graesser, Rus, & Hu, 2017; Woolf et al., 2009). The domain model contains the knowledge, skills, and strategies being tutored. It also includes ideal expert knowledge as well as bugs, mal-rules and student misconceptions. The student model contains a student's cognitive, motivational, affective and psychological states that are derived and tracked from their behaviors and performance during learning. The pedagogical model selects tutoring strategies, actions and steps on what the tutor should do next based on the domain model and student model. The user interface produces output in various media (e.g. texts, pictures, speech, sounds, animations, agents) after interpreting contributions via input media (e.g. typing, clicking, speech). These four components do not function in isolation. The student model enables ITSs to be adaptive by guiding the pedagogical model, which selects the strategies, actions, and steps in tutoring that are sensitive to the student's knowledge, skills, and abilities. VanLehn (2006) proposed two ways intelligent tutoring systems are configured to adapt to student model: the outer loop and the inner loop. The outer loop includes macro-adaptations whereas the inner loop includes micro-adaptations. The outer loop is the problem selection loop that decides what problem (e.g., task, main question, text to comprehend) to present next for the student to work on. The inner loop refers to the tutoring actions and steps within a problem that are sensitive to the student performance. ITSs aim at selecting learning materials within students' ZPD and choose strategies to guide students through outer loop and inner loop.

In order to build a student model that tracks knowledge states dynamically and more accurately, ITSs have adopted various theories and modeling methods. For instance, ALEKS mathematics tutor implemented a "knowledge space model" based on knowledge space theory (Doignon, & Falmagne, 1999). According to this theory, a knowledge state is the complete set of problems that an individual is capable of solving in a particular topic (e.g. Algebra). A domain model consists of a large number of possible knowledge states and a student model is a record of the knowledge states that are either mastered or not mastered by a student. The "inner fringe" refers to the items that differentiate a knowledge state from its immediate predecessor, and it also indicates the items that the student has mastered. The "outer fringe" refers to the items between a knowledge state and its immediate successor, and it indicates the items that a student is ready to learn. Specifically, outer fringe is what is within a student's ZPD. By identifying the inner fringe and outer fringe of a student's knowledge state, the knowledge space model decides what learning materials to provide to the student next (Doignon, & Falmagne, 1999; Falmagne, Albert, Doble, Eppstein, & Hu, 2013). Cognitive Tutor for mathematics developed by Carnegie Learning, utilizes a knowledge tracing model based on the ACT-R theory of human cognition (Anderson, 1996; Ritter, Anderson, Koedinger, & Corbett, 2007). The knowledge tracing model tracks the students' progress from problem to problem and builds a profile of their strengths and weaknesses compared to the domain model (Anderson, Corbett, Koedinger, & Pelletier, 1995). This profile is then used to determine when skills are learned. No new skills are introduced until the mastery of prior skills. This keeps the work within a student's ZPD by controlling the rate of progress in the Cognitive Tutor.

AutoTutor takes a different student modeling approach from ALEKS and Cognitive Tutor. AutoTutor is a conversation-based ITS that helps students learn by holding conversations in natural language (Graesser, 2016; Graesser et al., 2004). AutoTutor has been implemented in different systems across domains such as physics, computer literacy, and critical thinking. Regardless of the domain, AutoTutor dialogues are based on the expectation-misconception tailored (EMT) tutoring framework. For each lesson in AutoTutor there are pre-defined expectations and misconceptions provided by subject matter experts. The expectations consist of knowledge components (Koedinger, Corbett, & Perfetti, 2012) which are made of several key pieces of information needed to fully understand the lesson. The EMT framework evaluates the students' knowledge by comparing their answers with expectations and misconceptions relevant to the problems (Graesser, 2016; Graesser, Penumatsa, Ventura, Cai, & Hu, 2007). AutoTutor then scaffolds students through discourse moves (such as pumps and hints) that encourage the student to generate information. By using these various applications of student models, ITSs can dynamically evaluate a student's knowledge state, provide them with individualized learning materials, and guide them through the learning process with specific pedagogical strategies.

Applications of the ZPD can also be found in the interactive learning of ITSs. Chi and Wylie (2014) proposed the ICAP (Interactive, Constructive, Active, Passive) framework and defined four types of learning activities according to different modes of cognitive activities. The first type of learning activity, passive learning, is defined as, "being oriented toward and receiving information form the instructional material without overtly doing anything else related to learning" (Chi & Wylie, 2014, p. 221). An example of passive learning would be reading a text or listening to a lecture without doing anything else. The second type of learning activity is active learning, which refers to students conducting some form of motoric action or physical manipulation of the learning material. For example, underlining or highlighting while reading a text, or taking verbatim notes while listening to a lecture would be regarded as active learning. The third type of learning activity is constructive learning, which is defined as behaviors "... in which learners generate or produce additional externalized outputs or products beyond what was provided in the learning materials" (Chi & Wylie, 2014, p. 222). An example of constructive learning would be taking notes in one's own words while reading, or drawing concept maps while listening to a lecture. The last type of learning activity is interactive learning, which is defined as "…dialogues that meet two criteria: (a) both partners' utterances must be primarily constructive, and (b) a sufficient degree of turn taking must occur" (Chi & Wylie, 2014, p. 223). An example would be when a student asks and answers comprehension questions with a partner. According to the ICAP framework, students are more engaged and learn more in active learning than they are in passive learning, in constructive learning than they are in active learning, and in interactive learning than they are in constructive learning. One-on-one tutoring that involves interaction between a student and a human tutor is often considered among the most effective learning environment. The effectiveness of one-on-one tutoring is well documented, with effect sizes ranging from 0.4 to 2.0 sigma (Bloom, 1984; Cohen, Kulik & Kulik, 1982). Of the computer-assisted learning environments, ITSs are the closest technological analogue of one-to-one human tutoring (Durlach & Spain, 2014).

According to the theory of ZPD, scaffolding is a teaching method that helps a student successfully perform a task within his or her ZPD with the help a teacher or a more advanced student. In the learning environment of ITSs, computer agents play the role of a teacher or an advanced peer or both, interact with students by providing personalized instruction and feedback. ITSs vary in the way they provide feedback, and the granularity of the feedback. VanLehn (2011) distinguished the feedback provided by ITSs into answer-based, step-based, and substep-based based on the opportunities students get while interacting with ITSs when solving a problem. Answer-based ITSs do not provide feedback until the student completes a problem and submits a solution. If the answer is correct, the student may be presented with a new problem. If the answer is incorrect the student may get another chance to re-submit an answer or redo the task. Step-based ITSs provide guidance or feedback on each step of a solution, which means a student does not need to wait until the end of a problem to receive feedback. Substep-based ITSs provide the most detailed feedback by giving information or guidance before a student reach a correct solution of a step. For example, a computer agent may follow up a partially correct step asking the student a probing question and help the student get the correct answer. Answer-based, step-based, and substep-based ITSs all implement the element of scaffolding, but they differ from one another at the feedback granularity, which presumably affects the effectiveness of ITSs.

Aptitude-Treatment Interaction and Intelligent Tutoring Systems

The modern definition of aptitude-treatment interactions (ATI) stems from Cronbach (1957). Cronbach argued that cognitive treatments and the individual should be taken into consideration together, which would yield the best results because, "we can expect some attributes of a person to have strong interactions with treatment variables" (Cronbach, 1957, p. 680). Cronbach believed that a general intelligence test was too broad and tended to have little interaction with the treatment, and therefore could not properly guide differential treatments. He suggested designing treatments for groups of students with particular aptitude patterns, instead of designing treatments for the average person. He also suggested that the measure of aptitude need to be discovered.

A considerable amount of research applying the ATI method has observed interactions between student attributes and treatment variables. For example, Keislar and Stern (1970) conducted an experiment in which one group of third-graders was taught a single-hypothesis strategy, whereas the other group of third-graders was taught a multiple-hypothesis strategy. They found the high ability students performed better with the multiple-hypothesis strategy, but low ability students performed better with the single-hypothesis strategy. Skanes and his colleagues (1974) reported an experiment that explored the effect of practicing with a letter/ number series on learning. They reported students with lower IQ performed better with direct training on letters, while those with higher IQ performed better with indirect training using numbers. Cronbach did not specifically define “aptitude” in the ATI theory. Research that applied ATI regarded aptitude as mainly related to cognitive measures, such as prior knowledge, prior skill, or intelligence. Snow (1987) suggested that individual difference constructs needed greater consideration of the joint functioning between cognitive, conative, and affective processes. Snow (1991) reiterated this point and defined aptitude as a “...complex of personal characteristics identified before and during treatment that accounts for a person’s end state after a particular treatment” (Snow, 1991, p. 205). He also defined the domain of aptitude as, “not limited to intelligence or some fixed list of differential abilities but includes personality and motivational differences along with styles, attitudes and beliefs” (Snow, 1991, p. 205). Snow provided an example of a typical study of instruction where ATI could be applied. The study involved treatments that differed in structure and subjects with high or low cognitive ability. The high structured treatment was more didactic, direct, and teacher-centered and was predicted to be most successful with students of lower ability. Conversely, the low structured treatment was more inductive, indirect, and student-centered and was predicted to result in better learning for high ability students. Apart from ability, Snow suggested that anxiety could also be an aptitude measure. For highly anxious students, high structured treatment might work better than low structured treatment because it guided the students’ attention towards the tasks. Alternatively, he suggested that non-anxious students might do well with a less structured treatment because they do not require attention guidance.

In addition to the complexity of individual attributes, domain complexity should also be considered. Dance and Neufeld (1988) interpreted aptitude as a “trait” and assumed it to be stable, but Snow (1991) argued that aptitude should be a relational construct that included the mutual fit or unfit between a person and a situation. For example, a person good at math might not be good at reading comprehension. Snow (1991) emphasized that the characteristics of a situation should be included in the definition of aptitude. Snow (1989) developed a flow chart, shown in Figure 1, that outlined aptitude analysis, and how it can be applied. Overall, ATI methods emphasize the consideration of individual differences in treatment evaluations. The results of treatment(s) may differ for individuals with different measures of aptitude.

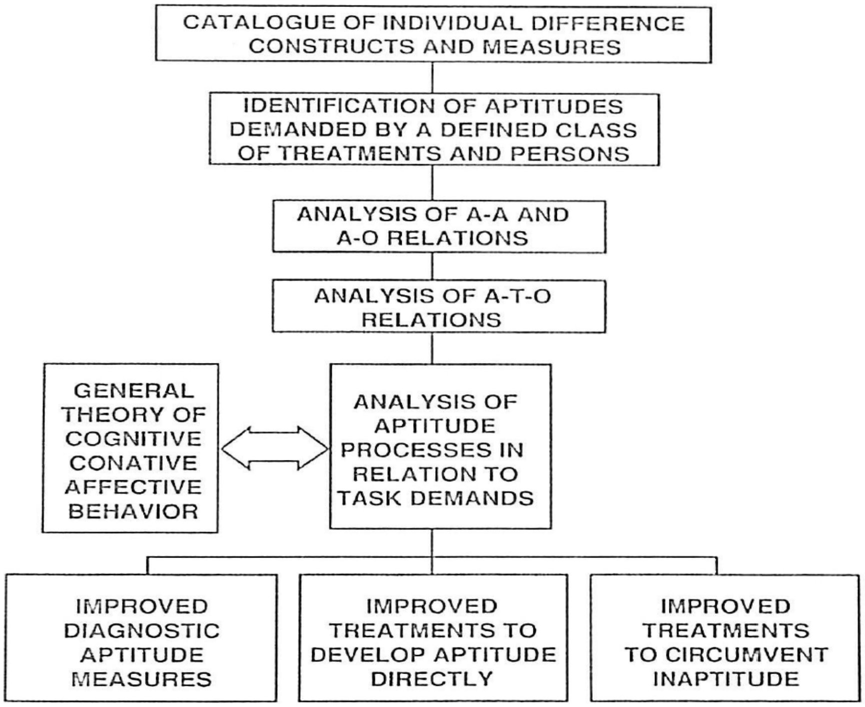

CATALOGUE OF INDIVIDUAL DIFFERENCE CONSTRUCTS AND MEASURES
IDENTIFICATION OF APTITUDES DEMANDED BY A DEFINED CLASS OF TREATMENTS AND PERSONS
ANALYSIS OF A-A AND A-O RELATIONS
ANALYSIS OF A-T-O RELATIONS
GENERAL THEORY OF COGNITIVE CONATIVE AFFECTIVE BEHAVIOR
ANALYSIS OF APTITUDE PROCESSES IN RELATION TO TASK DEMANDS
IMPROVED DIAGNOSTIC APTITUDE MEASURES
IMPROVED TREATMENTS TO DEVELOP APTITUDE DIRECTLY
IMPROVED TREATMENTS TO CIRCUMVENT INAPTITUDE

Figure 1: Snow's general approach to aptitude analysis.

(Adapted from Cognitive-conative aptitude in learning by Snow, R., 1989, In R. Kanfer, P.L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation and methodology: The Minnesota symposium on learning and individual differences* (pp. 435-474). Hillsdale, NJ: Erlbaum)

ATI methods have been frequently incorporated in ITSs research. Specifically, researchers have taken students' prior knowledge and skill into account as well as non-cognitive attributes (e.g., race and gender) when evaluating ITSs. For example, Beal, Walles, Arroyo, and

Wolf (2007) evaluated *Wayang Outpost*, an ITS for high school math for high school math and found that students with lower initial math skills showed greater improvement after interacting with the ITS than students with higher initial math skills. In another study, Beale at al. (2010) evaluated a different ITS for math, *AnimalWatch*, and found similar results. The low-skill students improved significantly after working with AnimalWatch, but the high-skill students did not. In the area of comprehension training, McNamara, O'Reilly, Best and Ozuru (2006) assessed the Interactive Strategy Training for Active Reading and Thinking (iSTART) system, , an ITS that was designed to teach reading comprehension strategies (McNamara, O'Reilly, Rowe, Boonthum, & Levinstein, 2007). They found after working with iSTART, students with lower prior knowledge performed significantly better on text-based questions, while students with higher prior knowledge improved significantly on bridge-inference questions.

Noncognitive attributes of students also may interact with learning with an ITS. For example, Smith (2001) studied the effect of the Carnegie Algebra Tutor and found that black students started with lower scores on the pretest than white students but had higher scores than white students on the posttest after working with the system. Craig et al. (2013) used an ITS called ALEKS to teach statistics and found a similar phenomenon; black students benefited more than white students from ALEKS. Specifically, the initial achievement gap between black and white students was eliminated after using ALEKS for one semester. It should be noted that this dissertation addresses cognitive attributes of students but not noncognitive attributes.

In addition to individual empirical studies, some meta-analyses included students' attributes as moderators while evaluating the effectiveness of ITSs, and reported different effect sizes of ITSs for different student groups. Ma et al. (2014) compared the effect of ITSs with traditional classroom teaching on students with different prior domain knowledge. They reported that the effect sizes of ITSs on students with low, medium and high prior knowledge were 0.37, 0.27 and 0.53, respectively. The meta-analysis performed by Steenbergen-Hu and Cooper (2013) restricted the domain of ITSs application to math and the school level of the students to k-12. This meta-analysis separated the studies including all levels of student from the studies only involved low achievers. They reported the effect of ITSs for helping general students (i.e., students at all levels) was greater than helping low achievers. Although the empirical studies and meta-analyses covered ITSs implemented in various domains, the majority of the ITSs were implemented in mathematics and science related domains.

Intelligent Tutoring Systems (ITSs) in the Domain of Reading Comprehension

Only a few of the studies in the previously mentioned ITS menta-analyses were in the reading comprehension domain. To better understand the effectiveness of ITSs in the domain of reading comprehension, a search for relevant studies was conducted in electronic databases including Web of Science, ERIC and PsycINFO and web search using Google and Google Scholar search engines. The following selection criteria were set up to guide the selection of studies:

(1) Evaluations included were empirical studies investigating the effect of intelligent tutoring systems or web-based learning system in the domain of reading comprehension.

(2) An independent comparison condition was in the experiment. The comparison condition could be traditional classroom, reading textbooks, or no treatment control. Studies with only one group using pretest-posttest design we excluded.

(3) Learning outcomes were measured quantitatively. Common measurements included standardized tests scores, locally developed test scores, and scores on tests developed by researchers.

(4) Studies had to use randomized experimental or quasi-experimental design and sufficient quantitative information was reported.

The key terms used in the search were the combination of *reading/reading comprehension* with one of the following terms: *intelligent tutoring, computer tutor, computer agents, web-based tutor and computer-assisted tutoring*. After the results were returned, the articles were compared with the inclusion criteria. There were 16 articles meeting the criteria. These studies evaluated 12 ITS or computer-assisted software. The ITSs and software include Interactive Strategy Trainer for Active Reading and Thinking, Intelligent Tutoring of the Structure Strategy, Althie's Alley, Tutoring With Alphie, Plato Focus, Academy of Reading, Destination Reading, the Waterford Early Reading Program, Headsprout, Leapfrog, Read 180, and Knowledgebox (Campuzano, Dynarski, Agodini, & Rall, 2009; Cheung, & Slavin, 2013; Chambers et al., 2008, 2011; Dynarski et al., 2007; Ji et al., 2018; McCarthy et al., 2018; McNamara et al., 2006; Meyer & Lei, 2017; O'Reilly et al., 2004; Wijekumar et al., 2012, 2013, 2017). The study samples included general students and struggling students, ranging from first grade to ninth grade. The students were either tutored in small groups or normal classrooms.

Two conflicting patterns were observed in these studies. First, one pattern indicated that educational software or ITSs were not different from traditional classroom teaching regarding their effectiveness. The other pattern indicated that the educational software was more effective than traditional classroom instruction.

Some of the studies examined the interaction between individual differences and learning technologies and the results were mixed. For example, Meyer and Lei (2017) investigated the gender difference and found larger learning gains for females than males from an ITS called Intelligent Tutoring of the Structure Strategy (ITSS). Wijekumar, Meyer and Lei (2017) also studied the interaction between gender and the ITSS, but they did not find a significant difference between males and females regarding learning gains. Meyer et al. (2010) and Ji et al. (2018) studied prior skill and found lower-skill readers benefited more from the ITTS system. Wijekumar, Meyer and Lei (2012) examined the effect of prior skill as well, but they did not find any difference between low-skill, medium-skill and high-skill readers in terms of their overall reading comprehension improvement. Meanwhile, they found high-skilled readers had larger learning gain than medium-skill and low-skill readers on one of the reading comprehension measures (i.e., main idea quality). O'Reilly, Sinclair and McNamara (2004) investigated the interaction between prior knowledge and an ITS called iSTART. They found students with less knowledge about interactive strategy performed significantly better on text-based questions while students with more knowledge about interactive strategy improved more on bridging–inference questions from iSTART training. Given the mixed results found in the relatively small number of studies, it remains unclear how ITSs interact with individual differences on learning.

Interaction between Individual Aptitude and Text Attributes

The Construction-Integration (CI) model (Kintsch, 1988, 1998) is considered one of the most complete models in discourse comprehension. According to the CI model, there are three levels of representations in comprehension: *the surface structure, the propositional textbase, and the situation model*. The surface structure represents the words and their syntactic relations. The propositional textbase refers to the underlying meaning of the explicit information in the text.

The situation model includes the references and inferences beyond the concepts explicitly mentioned in the texts. A good understanding of a text requires a reader to activate knowledge, integrate the knowledge into mental representations and build the connections between propositions. Guided by CI theory, Britton and Gulgoz (1991) found that improving text cohesion (i.e., overlap between sentences) led to better comprehension. McNamara et al. (1996) investigated the interaction between text cohesion and student's prior domain knowledge. They found that low-knowledge students learned from more coherent text. The low-knowledge students did not have the knowledge necessary to generate the inferences to connect separate ideas in the text, so they benefited more from text with high cohesion. Meanwhile, high-knowledge students benefited more from less coherent text because more knowledge was generated to bridge the gap when they read low-cohesion text. This effect was called the reverse cohesion effect.

Subsequent studies have considered reading skill when investigating the reverse cohesion effect. Voss and Silfies (1996) reported that the effect of text cohesion depended on reading skill. Their correlational analyses indicated that reading skill was primarily correlated with the better performance on high-cohesion texts, whereas prior knowledge correlated better with improved performance on low-cohesion texts. Linderholm et al. (2000) also examined reading skill but did not find any difference between high-skill readers and low-skill readers in terms of their benefit from high-cohesion texts.

More recent studies further investigated the interaction between reading skill, prior knowledge and text coherence (O'Reilly & McNamara, 2007; Ozuru et al., 2009). Both O'Reilly & McNamara (2007) and Ozuru et al. (2009) found only less skilled, high-knowledge readers benefited from low-cohesion texts, whereas the skilled high-knowledge readers benefited more from more coherent texts. The benefit of low-knowledge readers from the high-cohesion texts was mainly indicted in their response to inference questions instead of text-based questions (O'Reilly & McNamara, 2007). The degree that students benefited from more coherent texts was found to depend on their reading skill (O'Reilly & McNamara, 2007; Ozuru et al., 2009).

Task Difficulty and Engagement

According to the Goldilocks Principle, learning materials and tasks should not be too hard or too easy, but at the right level of difficulty for the students' skill level or prior knowledge (Graesser, 2009; Landauer & Psotka, 2000). Goldilocks Principle is in line with Vygotsky's zone of proximal development (ZPD), according to which the tasks should be too difficult for student to master on his/her own but can be achieved with the guidance and encouragement from a skilled partner (Vygotsky, 1978). It was found that individuals are challenged when they are given a task that demands skills, knowledge, or behaviors beyond their current capabilities (Van Velsor & McCauley, 2004). Additionally, individuals are motivated most by challenging tasks that provide an intermediate probability of success, instead of the ones that that offer certain success or failure (Belanich, Sibley, & Orvis, 2004; Malone & Lepper, 1987). As is shown in Figure 2, if a task is too difficult, individuals are likely to get anxious and give up; meanwhile, if a task is too easy, individuals are likely to get bored and quit.

Figure 2 illustrates the flow theory proposed by Csikszentmihalyi (1990). Flow is described as a positive state which engages students and elicits interest and absorption, and consequently sustains learning and motivation. Flow occurs when the task difficulty level and mastery skill are in balance (Csikszentmihalyi, 1990). If the task and the skill are out of balance, negative feelings may occur, such as anxiety, apathy, boredom, etc. (Csikszentmialyi, 1988, 1990; 1990; Massimini & Carli, 1988). Similar to engagement, task performance is the best when task difficulty is optimal (McShane & Travaglione, 2007). Performance improves when task difficulty increases, but the improvement drops if the tasks become too difficult or impossible. The relationship between task difficulty and task performance moves in the same direction as that between task difficulty and engagement. However, it is not clear whether affective states (e.g., anxiety, boredom) are associated with disengagement and influence performance negatively. For instance, D'Mello and Graesser (2012) found that students can be engaged without necessarily experiencing flow.

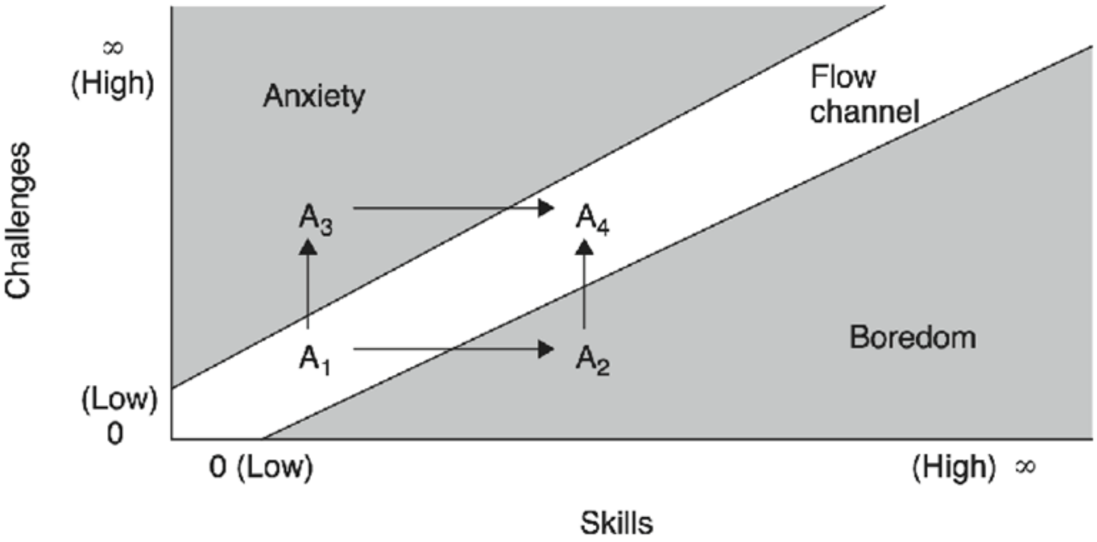


Figure 2. Relationship between challenge level and skill level.

(Adapted from *Flow: The psychology of optimal experience* (pp.74) by Czikszentmihalyi, M., 1990, New York, NY: Harper & Row.)

In educational settings, engagement is a multidimensional construct which has been used to describe behaviors, feelings, perceptions and attitudes (Reschly & Christenson, 2012). Fredricks, Bulmenfeld and Paris (2004) proposed three types of engagement: behavioral engagement, emotional engagement and cognitive engagement. Behavioral engagement is broadly defined as students' participation and involvement in learning tasks such as effort, persistence, attention, asking questions (Fredricks et al., 2004; Fredricks & McColskey, 2012). Emotional engagement refers to the affective reactions students experience during learning, such as interest, boredom, happiness, and anxiety (Fredricks et al., 2004; Skinner & Belmont, 1993). Cognitive engagement refers to students' psychological investment in learning tasks such as how they manage and control effort towards understanding and mastering knowledge and skills taught in school (Lamborn, Newmann, & Wehlage, 1992; Pintrich & De Groot, 1990). Pekrun and Linnenbrink-Garcia (2012) suggested there was some conceptual overlap between cognitive and behavioral engagement, and they further distinguished cognitive engagement (e.g. attention and memory processes) from cognitive–behavioral engagement (e.g., strategy use and self-regulation).

Regarding how to operationalize engagement, there was little consensus, and multiple measures were often recommended (Shernoff, 2013). For example, Fredricks et al. (2004) suggested engagement can either be self-reported or inferred by observations of behaviors. In this study, engagement was inferred from the learning behaviors when students interacted with AutoTutor. To be specific, the time students spend on reading texts (i.e., reading time) and answering questions (i.e., response time) were used to infer engagement. Regarding the amount of time students spent on a self-paced learning task, it is unclear if engagement should be reflected via faster (more concentration) or slower (more elaborative processing) reading and working times. Baker et al. (2008) suggested extremely short times or extremely long times are probabilistic signals of disengagement. Extremely short times signal that the reader was quickly perusing the material or "gaming the system" to get a correct answer without comprehending or learning. Extremely long times are a signal of mind-wandering (Feng, D'Mello, & Graesser, 2013; Smallwood, McSapadden, & Schooler, 2007) or simply taking a break and leaving the learning environment for a lengthy span of time. The current study adopted the concept "zone of engagement" (Greenberg, Graesser, Frijters, Lippert, & Talwar, 2018), and operationalized engagement as the proportion of behaviors within the zone of engagement. Specifically, it refers to the response times that were not too short or too long, but within a reasonable range. Another operational definition of engagement was the time of on-task behaviors (Shapiro, 2004), specifically the time students spent on reading texts and answering questions.

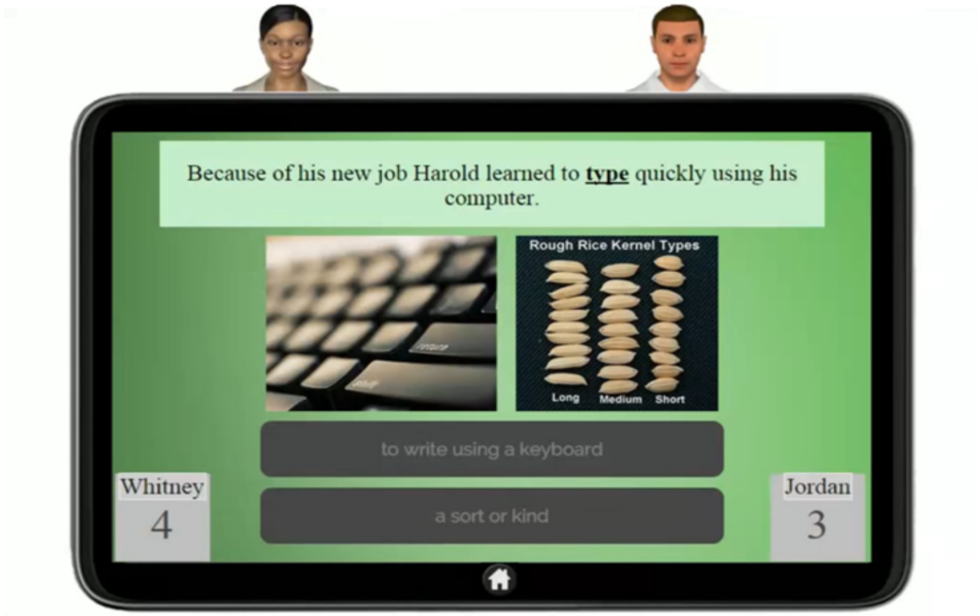
AutoTutor

AutoTutor is a conversation-based intelligent tutoring system (ITS) that has promoted learning on a wide range of topics (Graesser, 2016; Nye, Graesser, & Hu, 2014). AutoTutor has shown the average learning gains of 0.8 standard deviation units across topics, compared to traditional teaching comparison groups. AutoTutor holds a conversation with students applying an expectation-misconception tailored (EMT) approach (Graesser, Hu, & McNamara, 2005). A tutoring dialogue is made up of questions that assess a student's understanding of the content by comparing it to expected answers or misconceptions in real time. Using the EMT approach, AutoTutor is constantly assessing and helping the students by providing feedback, hints, pumps, and prompts to guide learning of the content.

Traditional AutoTutor systems implement conversations called *dialogues* that model the interactions that occur between a single human tutor and human student. More recent versions of AutoTutor utilize *trialogues* which are tutorial conversations between three actors: a teacher agent, a human student, and a peer agent (Graesser, Forsyth, & Lehman, 2017; Graesser, Li, & Forsyth, 2014). Trialogues have several advantages over dialogues. For example, in a trialogue design, the human student can model productive learning behaviors that are programmed into the peer agent. The peer agent may also express misconceptions that the human student shares and the negative feedback received from the tutor agent can be directed to the peer agent instead of the human student. This helps avoid many of the undesirable effects from receiving direct negative feedback. Trialogues also help students master more difficult learning material. For example, trialogues successfully helped students learn scientific reasoning skills in an AutoTutor offshoot called *Operation ARA* (Halpern et al., 2012; Mills, Graesser, Risko, & D'Mello, 2017).

Agent trialogues are implemented in *AutoTutor* for CSAL (Graesser et al., 2016), an ITS developed at the Center for the Study of Adult Literacy (CSAL). The web-based system is designed to help adults with low literacy acquire strategies for comprehending text at multiple levels of language and discourse. The system includes two computer agents (a teacher agent and a peer agent) which have conversations with human students and between themselves. The students are guided through their learning process by the computer agents. These three-way conversations are designed to (a) provide instruction on reading comprehension strategies, (b) help the student apply these strategies to particular texts, (c) assess the student's performance on applying these strategies, and (d) guide the student in using the digital facilities. While previous implementations of AutoTutor relied on written natural language input from the student, the students in AutoTutor for CSAL may have difficulties with writing. Thus, this version of AutoTutor was designed so that students interact through point-and-click, answering multiple choice questions, or using drag-and-drop. The conversational feature of AutoTutor still guides the student, but the questions can be solved without typed input.

The lessons typically start with a 2-3 minutes video that reviews a comprehension strategy. After the review, the computer agents present a text for them to read and hold a conversation with the student about the text. AutoTutor scaffolds students during learning by asking questions, providing short feedback, explaining how the answers are right or wrong, and filling in gaps of information. Figure 3 is an example of a "game mode" lesson in AutoTutor where the adult and peer agent compete to earn points by correctly answering questions about lesson material. The teacher agent (on the left) is asking both the student and the peer agent (on the right) to find out the meaning for the word "type" in the given context. The scores of both the student and peer agent are shown under their names. The student chooses the answer by clicking whereas the peer agent gives his answer by talking. Performance during learning with AutoTutor is measured by the accuracy of their answers to the questions and the response time to answer the question.



Because of his new job Harold learned to **type** quickly using his computer.

Figure 3: Example trialogue with competition which focuses on the meaning of words from context.

Text difficulty in AutoTutor lessons. Most AutoTutor lessons contain three texts which are of different difficulty level. The difficulty of the texts was calculated by a computer system called Coh-Metrix (Graesser, McNamara, & Kulikowich, 2011; Graesser et al., 2014). Coh-Metrix was developed to scale texts on dozens of components of language and discourse (http://cohmetrix.memphis.edu). Modules of Coh-Metrix use lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis and other components which are widely used in computational linguistics. The current public web site provides over 100 measures. The following are the five major dimensions of Coh-Metrix.

Narrativity. Narrative text tells a story, with characters, events, places, and things that are familiar to readers/listeners. Narrativity is closely affiliated with everyday oral conversation. This robust component is greatly affiliated with word familiarity, world knowledge, and oral language. Narrativity is contrasted by informational (or non-narrative) texts on less familiar topics.

Syntactic Simplicity. Syntactic simplicity reflects sentences with fewer words and in simpler, familiar syntactic structures which are comparatively easier to process and understand. Difficult sentences have more words positioned in complex, unfamiliar structures, which increase the difficulty of comprehension.

Word Concreteness. Texts are easier to process if they contain content words that are concrete and meaningful and evoke mental images and are more meaningful. Abstract words increase the difficulty to construct visual representations in the mind and make texts more challenging to understand.

Referential Cohesion. Texts with high referential cohesion contain words and ideas that overlap across sentences and the entire text, forming threads that connect the explicit textbase. Texts with low referential cohesion are more difficult to process and understand because there is information gap between sentences.

Deep Cohesion. Texts with causal and intentional connectives help the reader form a more coherent, explicit, and deeper understanding of the text at the level of the causal situation model. Ideas that are related semantically also contribute to deep cohesion. When texts contain many relationships but lack those connectives, inference is required to process the relationships between ideas in the texts.

Apart from the five major measures, Graesser et al. (2014) also identified a composite measure called *formality*, which can be used as a single approximate index of text difficulty. Coh-Metrix formality decreases with narrativity, syntactic simplicity, and word concreteness, but increases with referential cohesion and deep cohesion. The formula to calculate the formality score is as below.

Coh-Metrix Formality Score = (referential cohesion + deep cohesion - narrativity - syntactic simplicity - word concreteness)/5

Graesser et al. (2014) reported a 0.72 correlation between formality score and Flesch-Kincaid grade levels score (Klare, 1974) and a 0.66 correlation with Lexile scores (Stenner, 2006). Task difficulty in this dissertation was defined by the formality metric of Coh-Metrix.

Research Questions and Hypotheses

In summary, this dissertation draws its motivation from the zone of proximal development principle, the pursuit for ATI's, and empirical studies on the efficacy of ITSs. According to Vygotsky's zone of proximal development, learners benefit from tasks that are difficult but can be achieved with the guidance of a skilled partner (Vygotsky, 1978). This concept has often been implemented during the development of ITSs, but very little research has empirically investigated the validity of this concept, including the domain of reading comprehension. Individual differences need to be considered in order to assign tasks to individual students with the right level of difficulty. However, mixed results have emerged from previous research that specifically investigated the potential interaction between learner attributes and ITSs on learning in the domain of reading comprehension. Given these considerations, this dissertation investigated the following three questions: (1) Does task difficulty order (i.e., increasing difficulty versus decreasing difficulty) impact memory, engagement and performance during learning in a reading comprehension intervention with the AutoTutor ITS? (2) Do individual differences interact with task difficulty order in terms of their impact on memory, engagement, and performance during learning? If yes, how and to what extent do individual differences interact with task difficulty order? (3) Does engagement moderate the effect of task difficulty order on memory and performance during learning?

A within-subjects experiment was conducted that had two conditions: (1) an increasing difficulty order condition in which easy learning tasks were followed by difficult tasks versus (2) a decreasing difficulty order condition in which difficult learning tasks were presented before easy tasks. Task difficulty was defined according to the difficulty of the texts that students read, as measured by Coh-Metrix. Students were guided through the learning tasks, including reading texts and answering questions asked by computer agents in a conversation-based ITS, AutoTutor. The students' learning performance (i.e., accuracy and time to answer questions) was recorded during their interactions with AutoTutor. The students' reading skills were assessed by a reading comprehension test whereas memory was assessed by a recognition test on content delivered by AutoTutor. Engagement during learning was inferred from time data recorded by AutoTutor.

Hypothesis for Students with Low Reading Skills. A hypothesis based on ZPD and flow theory expects the low-skill readers to benefit more from the increasing difficulty order condition than the decreasing difficulty order condition. Schell's flow channel (2014), which illustrates the relationship between skill level and challenge level can help us interpret this hypothesis. As is shown in Figure 2, when the tasks become more difficult in the increasing difficulty condition, the challenges for low-skill students are expected to be at A1 and A3.

However, the actual challenge from the difficult tasks will be at point A4 due to prior learning experience. The challenge level versus skill level should keep the low-skill students in the flow channel throughout the whole learning process in the increasing difficulty order condition. On the other hand, in the decreasing difficulty condition the low-skill students will start with tasks beyond their skill level, which was A3. When the tasks are outside of their ZPD, the low-skill students will be anxious and will only stay in the flow channel only during the easy tasks. Therefore, the increasing difficulty order condition should benefit low-skill readings more in terms of their memory, engagement, and performance during learning.

Hypothesis for Students with High Reading Skills. There were two competing hypotheses for high-skill students. One hypothesis is there will be no difference between the increasing and decreasing difficulty order condition. As is shown in Figure 2, the challenge from easy tasks and difficult tasks can be represented by A2 and A4. Here, high-skill students are expected to get bored during easy tasks and move into the flow channel during difficult tasks. The effective learning will only happen at A4, which will be the time the high-skill students are engaged. The engaged and effective learning experience for high-skill students should only be associated with the more difficulty learning tasks. As a result, the high-skill students' memory, engagement and performance during learning should not be affected by the manipulation of task difficulty order. The alternative hypothesis for high-skill students expects the decreasing difficulty order condition to benefit students more than the increasing difficulty order condition in terms of their effect on memory, engagement and performance during learning. High-skill students are expected to get bored during their interactions with AutoTutor when they are provided with learning tasks that are too easy at the beginning. Boredom may then be carried over to the subsequent learning stage with when they are provided with difficult learning tasks.

Give this situation, the high-skill students will stay at point A2 for most of the learning process in the increasing difficulty order condition. In the decreasing difficulty order condition, the high-skill students will be given difficult tasks from the start, which will be challenging for them and can get them into the flow channel. If this affective state is carried over to the following learning stage when they are provided with easy tasks, the high-skill students should remain at point A4 for most of the learning process in the decreasing difficulty order condition. As a result, the high-skill students should have better memory, engagement, and performance during learning in the decreasing difficult order condition compared to the increasing difficulty order condition.

Chapter 3 Methods

Participants

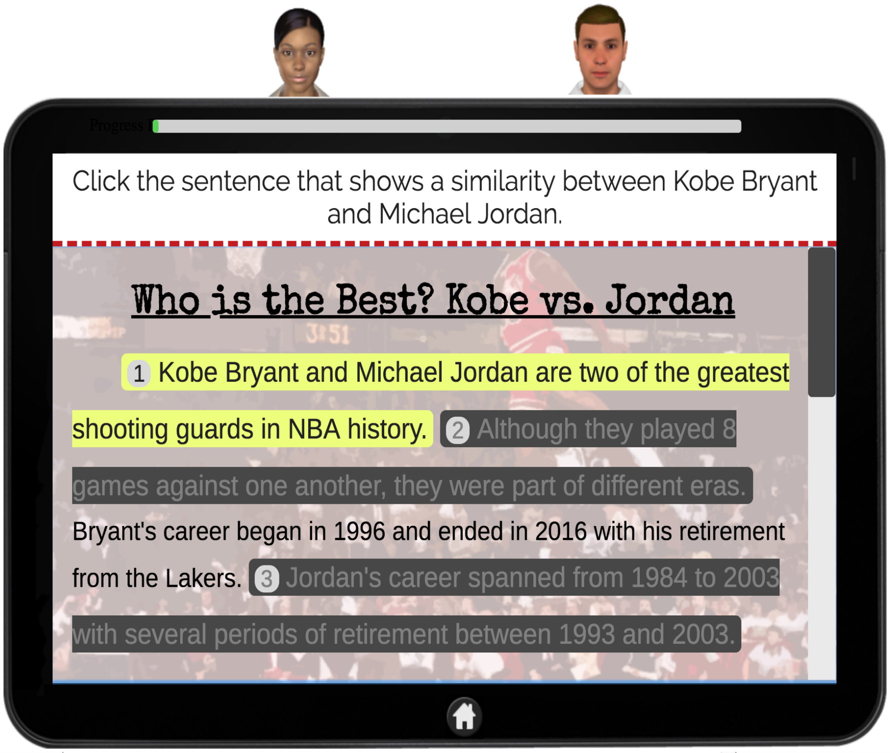
There were 159 participants recruited from the University of Memphis subject pool. The majority of the participants were female (60.4%) and native English speakers (87.4%). Race and ethnicity were reported as follows: the majority of the sample were White (43.4%), followed by Black/African American (42.8%), Hispanic/Latino (6.3%), Asian (4.4%) and Multicultural (3.1%). Participants ranged in age from 18 to 48 with a mean age of 21 (SD = 4.68).

Materials

AutoTutor. AutoTutor was used as the learning environment where the participant students interacted with two computer agents named Cristina (teacher agent) and Jordan (peer agent). Cristina is an African American female teacher agent, and Jordan is a racially ambiguous Latino male peer agent. Each student selected a nickname from a given list after logging into the system, which the computer agents used to refer to the student during the trialogues. The screenshot in Figure 4 illustrates the interface in the experiment.

Each lesson started with an introduction of the comprehension strategies highlighted in the lesson and then a text was presented to students. After students finished reading the text, they could click the “continue” bar and would then receive questions associated with the text. These questions were embedded in the trialogue conversations among the two agents and student. Each question and its corresponding answers were presented on the same screen. When a question was shown on the screen, the teacher agent (i.e., Cristina) asked students a question by reading it or paraphrasing it. The teacher agent asked both human students and the peer agent to answer the questions. Sometimes the teacher agent asked the human students to answer first, and sometimes the teacher agent asked the peer agent to answer first. The trialogue for a question was not fixed.

It depended on which choice the student and the peer agent selected. If a student and the peer agent were both correct, the teacher agent would tell them they were right, and explain why that answer was correct. If a student or the peer agent selected an incorrect answer, the teacher agent would first tell them who was correct and who was wrong, and then explained why one answer was wrong and the other was correct. If a student and the peer agent were both wrong, the teacher agent would tell them they were both wrong, and then explain why the answer they selected was wrong. During the tutoring session, the computer agents guided the students through the learning process by asking questions, providing short feedback, and explaining how the answers were right or wrong. The following is an example of a trialogue among Cristina (teacher agent), Jordan (peer agent) and the user who selected the nickname "Sam".



Click the sentence that shows a similarity between Kobe Bryant and Michael Jordan.

Who is the Best? Kobe vs. Jordan

1 Kobe Bryant and Michael Jordan are two of the greatest shooting guards in NBA history. 2 Although they played 8 games against one another, they were part of different eras. Bryant's career began in 1996 and ended in 2016 with his retirement from the Lakers. 3 Jordan's career spanned from 1984 to 2003 with several periods of retirement between 1993 and 2003.

Figure 4. An example screenshot of the learning environment where students interact with the computer agents.

Cristina: Sam and Jordan, now that we have read the passage, let's compare the careers of Michael Jordan and Kobe Bryant.

Gordan: I like Kobe more, even though my name is Jordan too!

Cristina: Oh! You're too much! Let's see if we can find some similarities and differences between them.

Cristina: Sam, click the sentence that shows how Kobe Bryant and Michael Jordan are similar. (Suppose the user selects "Kobe Bryant and Michael Jordan are two of the greatest shooting guards in NBA history")

Cristina: Jordan, what is your answer?

Gordan: Cristina, that is what I would choose. We are in agreement!

Cristina: Sam and Jordan, nice work! You are both right!

Cristina: Signal words can help us to see the comparison. In the first sentence, the use of the word "two" signals they have that information in common. In sentence two, the word "different" indicates a contrast.

Gordan: The first sentence says Kobe and Jordan are two of the greatest shooting guards. I think this is a similarity between them.

Cristina: The first sentence shows one of their similarities. They are two of the best shooting guards in basketball history.

The students interacted with AutorTutor on two lessons. The topics of the two lessons were "Compare and Contrast" and "Inferences from Texts". There were two texts associated with each lesson, one being relatively difficult and the other being easier. In the "Compare and Contrast" lesson, the formality scores of the difficult and easy texts were 0.19 and -0.31, respectively, on a z-score scale that was normed on a sample of over 37,000 texts (Graesser et al., 2014). The Flesch-Kincaid grade level of the two texts were 8.3 and 4.0, respectively. In the "Inferences from Texts" lesson, the formality scores of the two texts were 0.15 and -0.32. The Flesch-Kincaid grade level of the two texts were 8.9 and 5.4. There were 9-10 questions associated with each text.

The design of the two lessons was slightly different. In the Compare and Contrast lesson, students were allowed two attempts for each question. In other words, if a student did not answer a question correctly on the first try, the teacher agent would ask him/her to try it again after telling the student the answer was incorrect without explaining why. After the second attempt, the teacher agent would provide the correct answer and provide a detailed explanation. In the Inferences from Texts lesson, students were only allowed one attempt. This means the teacher agent would provide the correct answers with detailed explanations after the first attempt, and a student would only have one chance to answer a question. Given the design difference between the two lessons, later analyses only included first attempt performance measures in AutoTutor.

AutoTutor measures. Students' interaction with the computer agents were recorded turn by turn. The system behavior and user behavior were both recorded for each turn (e.g. system loading a page, student submitting an answer). The time (i.e., starting time, ending time, duration) associated with a behavior was also recorded. For instance, when a student selected an answer for a multiple choice question, the question, the alternative answers, and the answer selected by the student were recorded. Additionally, the conversation between computer agents and the student, the time the student spent on answering the question (i.e., from the time a question was presented on the screen to the time a student clicked an answer), and whether the answer was correct or incorrect were recorded by the system. With detailed records containing rich information about the learning process, we were able to explore various learning behaviors without observing the students in person. This study focused on three measures recorded by AutoTutor: reading time, response time, and performance accuracy. Reading time refers to the time students spent on reading a text. It was measured from the onset of an article page (i.e., the article was shown on the computer screen) to the time students clicked the “continue” button to move to the next page. Response time refers to the time students spent answering a question. It was measured from the onset of a question (i.e., a question shown on the computer screen) to the click on an option indicating the student’s answer. Both reading time and response time were recorded in milliseconds. Performance accuracy referred to the proportion of correctly answered questions. The result of each attempt on a question was measured as being correct or incorrect. Performance accuracy was computed using the number of correct answers divided by the number of total answers from a student.

Maze subtest. Reading skill assessment was administered to students on a webpage using the maze technique developed by Educational Testing Service (ETS) (Sabatini et al., 2019). The maze technique uses a forced-choice cloze paradigm and was found to be an indicator of basic reading efficiency and comprehension in several previous studies (Sabatini et al., 2019; Wayman et al., 2007). In a maze subtest, students were presented with three passages that were embedded with fill-in-the-blank sentences. There were three choices for each blank and students selected one of them to complete the sentence. Each passage had a time limit of 3 minutes. A short example is: At the earth’s core, the temperature is very (high/smile/twenty). Molten rock and hot gases sometimes surge toward the (jump/surface/plenty). When this happens, a volcano may (type/blanket/erupt). The raw score of the maze subtest was measured as the total number of correct answers. Time on the test was measured from the onset of the first sentence in the first passage to the time students finished the last sentence in the third passage and clicked the arrow to continue. A scaled score was generated based on the raw score and each individual's pattern of items answered correctly according to item response theory (IRT) (Sabatini et al., 2019). The scaled score considered the difficulty of an item as well as gave more weight to items that correlated higher with the raw score. As such, two students with the same raw score but different scaled scores answered different items correctly; the student with the higher scaled score answered items with better correspondence to higher ability levels.

Two fluency measures (i.e., words per minute, correct words per minute) were generated based on the characteristics of the reading materials and students' response time. Each maze subtest sentence was treated as an item while creating the fluency measures. The response time of an item (i.e., sentence) was measured from the onset of the sentence (i.e., the sentence is shown on the computer screen) until the time students clicked on the arrow to continue to next sentence. For each item, word per minute (WPM) was calculated using the sentence length (i.e., number of words in sentence) divided by response time in seconds multiplied by sixty. Each student read many sentences in the maze subtest, and the median WPM across all the sentences was generated as the WPM indicator for the student. The measure for words correct per minute (WCPM) was generated using the product of WPM and a student's raw accuracy (raw score divided by number of items attempted).

Recognition test. Students were given a recognition test that consisted of four alternative forced-choice questions at the end of the experiment. The recognition test assessed the students' memory for surface structure (i.e., wording and syntax) and meaning (Graesser & Mandler, 1975). The differentiation between surface structure and meaning was based on the "multistore" theories of memory, which suggest that phonemic, surface level information is stored in short-term memory whereas semantic, abstract information is store in long-term memory (Baddeley, 1966). The surface information is supposed to be lost quickly, while meaning is believed to be maintained longer (Kintsch, 1970; Wickelgren, 1972). In the recognition test, each question asked the student to identify the verbatim sentence that appeared in the texts they read (i.e., same surface structure and meaning, S+M+). The other three choices included one sentence that had the same surface structure but the meaning of sentence was altered (i.e., S+M-); one sentence that altered the surface structure (e.g., word order or syntax were changed) but the same meaning was kept (i.e., S-M+); and one sentence had both different surface structure and meaning (i.e., S-M-). The selection of S+M+ or S-M+ was regarded as correctly recognizing the meaning of the sentence from the text (i.e., memory for surface structure). The selection of S+M+ or S+M- was regarded as correctly recognizing the surface structure of the sentence from the text (i.e., memory for meaning). There were three questions for each text and students received a total of twelve questions for the four texts they read in the two AutoTutor lessons. Appendix D presents the items used in the recognition test. The score of recognized surface structure and meaning for each question was computed separately. The proportion of correctly recognized surface structure and correctly recognized meaning were used as the measures of recognition memory.

Design and Procedure

The experiment used a within-subjects design. Students interacted with conversational computer agents in a learning environment called AutoTutor (Graesser et al., 2016), which was introduced in Chapter 1. Each student was provided with two AutoTutor lessons, and each lesson had two texts with different difficulty. There were nine to ten multiple choice questions associated with each text. The text and associated questions were arranged in two contrasting orders: increasing difficulty order and decreasing difficulty order. Specifically, increasing difficulty means the easier text and associated questions were presented first, followed by the more difficult text and questions. Decreasing difficult order means the more difficult text and associated questions were presented before the easier text and associated questions. Each lesson was organized in a decreasing difficulty order as well as an increasing difficulty order, and the students received the combination of two lessons in two contrast difficulty orders. In other words, the students were randomly assigned one of the four sequences in Table 1. AutoTutor recorded students' learning behaviors during their interaction with computer agents. The records included subject information (i.e., subject system ID), events (e.g. page loading, a conversation turn starting), times (e.g. staring time, duration), actions (e.g. reading, answering questions), and results (e.g. whether an answer was correct).

Table 1
Lesson combinations in AutoTutor

Sequence	Lesson 1 - Order	Lesson 2 - Order
1	Compare and Contrast - Increasing	Inferences from Texts - Decreasing
2	Compare and Contrast - Decreasing	Inferences from Texts - Increasing
3	Inferences from Texts - Decreasing	Compare and Contrast - Increasing
4	Inferences from Texts - Increasing	Compare and Contrast - Decreasing

Notes. Increasing = increasing difficulty order; Decreasing = decreasing difficulty order.

Before the experiment started, students were given a consent form informing them the purpose of the study, the procedures they would experience, and the potential risks and benefits of participation (see Appendices A and B). After agreeing to participate in the study, students were given a short demographic survey with 5 questions, which asked the students their age, gender, ethnicity, native language and how many years they have learned English (see Appendix C). Next, students were given a link which directed them to an online reading assessment, the maze subtest. Students used the unique accounts and passwords provided to them to log in and work on the test. After logging in, students were presented with three passages with embedded fill-in-the-blank sentences, each with three choices. Students picked which of the three words correctly completed the sentence. Each passage had a time limit of 3 minutes. If a student finished a passage in less than 3 minutes, they could immediately move to the next one. If they did not finish the passage within 3 minutes, the rest of the sentences were not shown. Instead, the sentences in the next passage were presented on the screen.

After the maze subtest, students were directed to AutoTutor to start their interaction with the computer agents. The students were instructed to select a nickname from a name table to use in the system and then started to interact with the two computer agents (i.e., teacher agent and peer agent). For each lesson, the students were first provided with a text to read. When they finished reading, they could click the "continue" button to proceed to the AutoTutor conversational interaction that included the comprehension questions. Each question had three multiple choice options. The students clicked on their choice and then the teacher agent would ask the peer agent what his choice was. Sometimes the teacher agent asked the peer agent to answer the question first, and then let the student give answer. The peer agent sometimes agreed with the student, and sometimes selected a different answer. After the peer agent provided his choice, the teacher agent gave feedback on the correct answer and why it was correct. Sometimes the teacher agent commented on the student's or peer agent's incorrect answers. When the students finished all the questions following a text, they were provided another text and the corresponding questions. After they finished the two texts and associated questions, one lesson was completed. The students were then given the other lesson and repeated the same procedure.

The last phase of the experiment was a recognition test. Students were asked to select the sentences explicitly presented in the text or spoken by the agent. Each item had four alternatives of surface structure (same versus different) and meaning (same versus different). There were a total of twelve questions, with three questions from each text (see Appendix D).

The consent form, demographic survey and recognition test were completed on Qualtrics, an online software for conducting surveys and recording survey data. The maze subtest and AutoTutor lessons were completed on webpages.

Chapter 4 Results

In this chapter, the results of statistical analyses are reported to address the three research questions. In the first section, the effect of task difficulty order on recognition memory was examined by repeated measures ANOVA tests. In the second section, the interaction between task difficulty order and reading skill was investigated. Cluster analysis was performed to group student at different reading skill levels. Mixed-effects models were conducted to test the interaction between task difficulty order and reading skill level in terms of their effect on memory, engagement and performance during learning. In the last section, the moderation effect of engagement on memory and performance during learning was tested by mixed-effects models.

Effect of Task Difficulty Order on Recognition Memory

To investigate the effect of task difficulty order on student's recognition memory, analyses were conducted to compare student's recognition memory in the increasing difficulty order condition with the decreasing difficulty order condition.

There were four alternative answers for each question in the recognition test. The design of the alternative answers was: same surface structure and same meaning (i.e., S+M+), same surface structure and different meaning (i.e., S+M-), different surface structure and same meaning (i.e., S-M+), different surface structure and different meaning (i.e., S-M-), compared to the verbatim sentence (i.e., S+M+) that appeared in one of the texts. The selection of S+M+ or S-M+ was regarded as memorizing the meaning of the sentence from the text. The selection of S+M+ or S+M- was regarded as memorizing the surface structure of the sentence from the text. The descriptive statistics (i.e., means and standard deviations) of the students' performance (i.e., memory for surface structure and meaning) on the recognition test in two task difficulty order conditions are shown in Table 2.

Table 2

Means and standard deviation of memory for surface structure and meaning

	Surface Structure	Meaning
Decreasing Difficulty	0.70 (0.21)	0.68 (0.20)
Increasing Difficulty	0.71 (0.18)	0.67 (0.19)

Repeated measure ANOVAs were performed to compare the memory for surface structure and meaning in decreasing difficulty order condition with that of increasing difficulty order condition. There was no statistically significant difference in students' memory for surface structure between the two task difficulty order conditions, F (1,157) = 0.27, p = 0.602. There was no statistically significant difference in students' memory for meaning between the decreasing difficulty order condition and the increasing difficulty order condition, either, F (1,157) = 0.16, p = 0.689.

Table 3

Means and standard deviations of memory for surface structure and meaning in two task conditions for three types of questions

	Surface Structure			Meaning		
	Main	Support	Queried	Main	Support	Queried
Decreasing	0.82 (0.31)	0.70 (0.29)	0.65 (0.36)	0.63 (0.43)	0.69 (0.28)	0.66 (0.34)
Increasing	0.84 (0.30)	0.72 (0.29)	0.61 (0.35)	0.63 (0.42)	0.68 (0.28)	0.66 (0.30)

Notes. Decreasing = decreasing task difficulty order, Increasing = increasing task difficulty order, Main = main idea questions, Support = supporting detail questions, Queried = questions asking about information queried by computer agents.

The questions in the recognition test were further categorized into three types. They were questions asking about the main idea of the text, the supporting details in the text, or the information which was queried by the computer agents when students interacted with AutoTutor. The students' memory associated with each type of questions was examined. Table 3 shows the descriptive statistics (i.e., means and standard deviations) of students' memory for surface structure and meaning segregated for the three types of questions in the decreasing versus increasing difficulty order conditions.

Repeated measure ANOVAs were conducted on each type of question to compare students' memory for surface structure or meaning in decreasing difficulty order condition versus increasing difficulty condition. The results indicated that there was no statistically significant difference in the memory for surface structure between decreasing difficulty order condition and increasing difficulty order condition for the main idea questions (F (1,157) = 0.43, p = 0.516), supporting detail questions (F (1,157) = 0.21, p = 0.644) and the queried questions (F (1,157) = 0.88, p = 0.350). There was no statistically significant difference in students' memory for meaning between the two task difficulty order conditions for the main idea questions (F (1,157) = 0.01, p = 0.905), supporting detail questions (F (1,157) = 0.33, , p = 0.569) and queried questions (F (1,157) = 0.03, p = 0.866).

Interaction between Task Difficulty Order and Reading Skill

In this section, statistical analyses were conducted to investigate the interaction between task difficulty order and reading skill in terms of their effect on recognition memory, engagement, and performance during learning.

The students' reading skills were measured by the maze subtest, and the outcome measures of the test were used to differentiate students. Four outcome measures were generated by the maze subtest: raw score, scaled score, words per minute and words correct per minute. The correlation matrix of the four measures is shown in Table 4.

As is shown in Table 4, the raw score and the scaled score were highly correlated, r (156) = 0.85, p < .01. Words per minute and words correct per minutes were nearly perfectly correlated, *r* (156) = 0.99, *p* < .01. The distributions of the four measures were further examined. The density plots of the four variables are shown in Figure 5, 6, 7 and 8. As is indicated by the plots, the distributions of raw score and scaled score are left skewed. The skewness and kurtosis of raw score and scaled score are: skewness$_{\text{raw score}}$ = -3.83, kurtosis$_{\text{raw score}}$ = 20.46, skewness$_{\text{scaled score}}$ = -1.18, kurtosis$_{\text{scaled score}}$ = 4.02. The distributions of words per minute and correct words per minute are right skewed. The skewness and kurtosis of words per minute and correct words per minute are skewness$_{\text{words per minute}}$ = 1.60, kurtosis$_{\text{words per minute}}$ = 9.73, skewness$_{\text{words correct per minute}}$ = 1.38, kurtosis$_{\text{words correct per minute}}$ = 9.46. The distributions of words per minute and words correct per minutewere similar. Considering the correlations and distributions the four measures, scaled score and words per minute were selected to differentiate students' reading skill level.

Table 4
Correlations between the four maze subtest measures

		Raw Score	Scaled Score	WPM	WCPM
Raw Score	Correlation Coefficient				
	Sig. (2-tailed)	—			
	N				
Scaled Score	Correlation Coefficient	0.85**			
	Sig. (2-tailed)	.000	—		
	N	158			
WPM	Correlation Coefficient	0.43**	0.45**		
	Sig. (2-tailed)	.000	.000	—	
	N	158	158		
WCPM	Correlation Coefficient	0.50**	0.53**	0.99**	
	Sig. (2-tailed)	.000	.000	.000	—
	N	158	158	158	_

Notes. WPC = words per minute, WCPM = words correct per minute, **Correlation is significant at the .01 level (2-tailed).

Cluster Analysis. Cluster analysis is a statistical exploratory tool used to find similar groups in an unsupervised fashion. It partitions objects into clusters so that the objects in the same cluster are more similar to each other than to those in other clusters. A cluster analysis based on students' scaled score and words per minute was performed in order to group the students based on their reading skills. A k-means clustering algorithm was first applied to the data. K-means clustering fits data points into clusters by iteratively reassigning and re-averaging the cluster centers until the points have reached convergence (Hartigan & Wong, 1979; Jain, 2010). It is a common choice for clustering data since it is simple, effective and relatively efficient. R (version 3.3.4) was used to perform the k-means clustering algorithm of Hartigan and Wong (1979). The disadvantage of k-means clustering is that it is sensitive to the initial centroids and also does not do well with clusters with non-spherical shape and different size (Jain, 2010). In light of this, a hierarchical cluster analysis was conducted as well. Hierarchical clustering is different from k-means clustering in that it directly divides a dataset into a number of disjoint groups. It proceeds successively either by merging smaller clusters into larger ones (bottom-up),

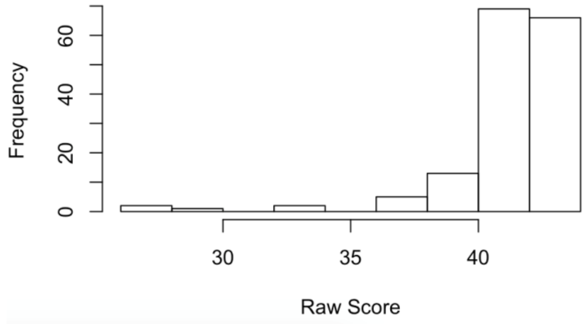


Figure 5: Distribution of raw score.

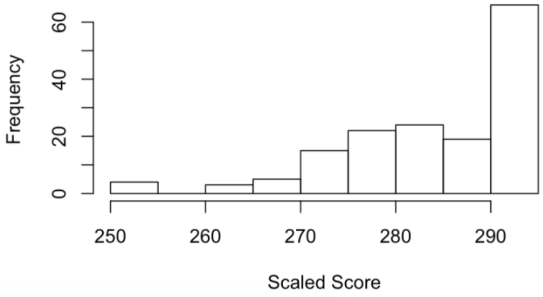


Figure 6: Distribution of scaled score.

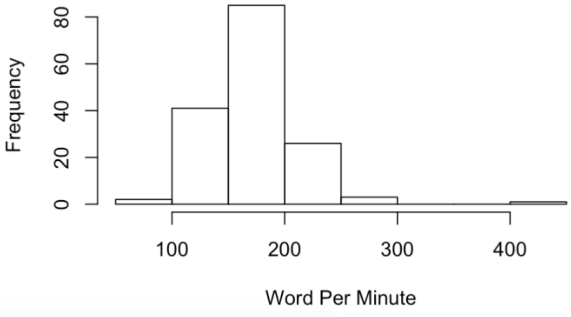


Figure 7: Distribution of words per minute.

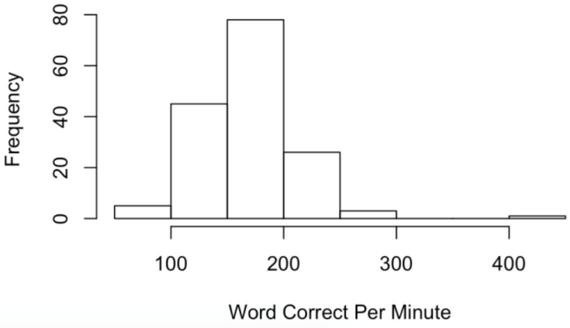


Figure 8: Distribution of correct words per minute.

or by splitting larger clusters into smaller clusters (top-down) (Jain, Murty, & Flynn, 1999). A hierarchical clustering was performed using Ward's method (Ward Jr, 1963) on the data. To select the optimum clustering method and number of clusters, the 2-cluster, 3-cluster and 4-cluster solutions using k-means clustering algorithm were compared with the 2-cluster, 3-cluster and 4-cluster solutions using hierarchical clustering algorithm. The R package used for the comparisons was clValid (Brock, Pihur, Datta, & Datta, 2008). The scores of the six solutions were computed on three measures, namely connectivity, Silhouette Width, and Dunn Index. Connectivity measures the degree of connectedness of the clusters, and Silhouette Width and the Dunn Index measure compactness and separation of the clusters. The comparison of the six solutions on the three measures indicated that the best solution was the 2-cluster solution using hierarchical clustering algorithm. The final result of the hierarchical clustering algorithm is a binary tree of clusters called dendrogram, which shows how the clusters are related to each other. Figure 9 shows the dendrogram of hierarchical clustering using Ward's method (Ward Jr., 1963).

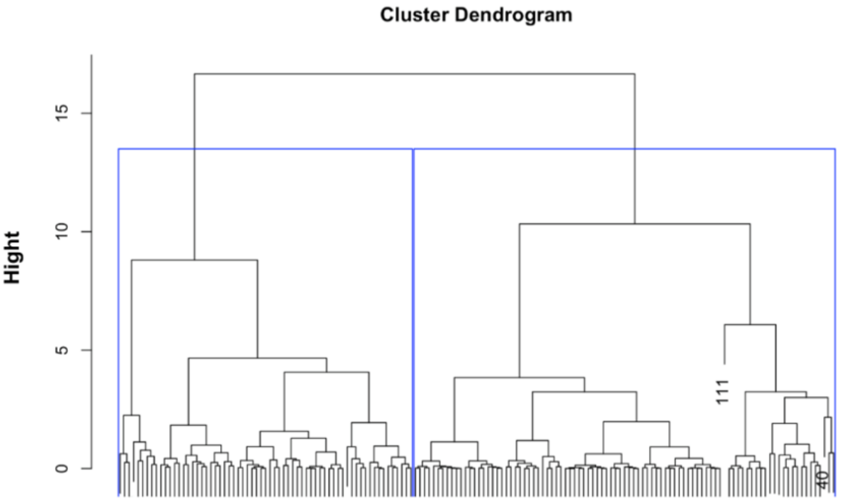


Figure 9: Dendrogram from hierarchical clustering using Ward's method.

Cluster 1 had the average scaled score of 276.26 (*SD* = 8.68) and average reading speed of 147.23 (*SD* = 24.19) words per minute. Cluster 2 had the average scaled score of 291.49 (*SD* = 4.53), and average reading speed of 194.22 (*SD* = 39.63) words per minute. The reading score and reading speed of cluster 1 was lower than that of cluster 2, so cluster 1 was labeled as low-skill readers and cluster 2 was labeled as high-skill readers. Table 5 presents the descriptive statistics (i.e., means and standard deviations) of the memory for surface structure and meaning associated with low-skill readers and high-skill readers in the two task difficulty order conditions (i.e., increasing difficulty order condition and decreasing difficulty order condition).

Table 5
Means and standard deviations of high-skill and low-skill readers' memory for surface *structure and meaning in two task conditions*

	Condition	*n*	Surface Structure	Meaning
High Skill	Decreasing	93	0.73 (0.21)	0.68 (0.21)
	Increasing	93	0.70 (0.18)	0.66 (0.19)
Low Skill	Decreasing	65	0.65 (0.20)	0.67 (0.20)
	Increasing	65	0.72 (0.17)	0.67 (0.18)

Notes. Decreasing = decreasing task difficulty order condition, Increasing = increasing task difficulty order condition.

A mixed-effects linear regression was performed to predict the memory for surface structure as a function of reading skill level, task difficulty order condition, and their interaction. Subjects were specified as a random-effect factor to adjust for the subject variance. The results were statistically nonsignificant for the main effect of task difficulty order condition (*t* (156) = 2.63, *p* = 0.010), and the main effect of reading skill level (*t* (305) = 0.62, *p* = 0.538). The interaction between task difficulty order condition and reading skill level was statistically significant (*t* (156) = 2.58, *p* = 0.011). Follow-up planned comparisons were carried out using repeated measure ANOVAs to compare the memory for surface structure in two task difficulty order conditions for both high-skill readers and low-skill readers. Results indicated that high-skill readers' memory for surface structure between two conditions was not statistically significant, *F* (1, 92) = 1.49, *p* = 0.225. Low-skill readers' memory for surface structure between two conditions was statistically significant, *F* (1, 64) = 5.77, *p* = 0.019. Low-skill readers had better memory for surface structure in the increasing difficulty order condition compared to decreasing difficulty order condition.

As in the analysis of memory for surface structure, a mixed-effects linear regression was conducted on memory for meaning. Results indicated that there was no statistically significant interaction between task difficulty order condition and reading skill level (*t* (156) = 0.33, *p* = 0.739). The results were not statistically significant for the main effect of task difficulty (*t* (156) = 0.20, *p* = 0.841), or the main effect of reading skill level (*t* (309) = 0.30, *p* = 0.766). The high-skill and low-skill readers' memory for surface structure and meaning in two task difficulty order conditions are shown in Figure 10 and Figure 11.

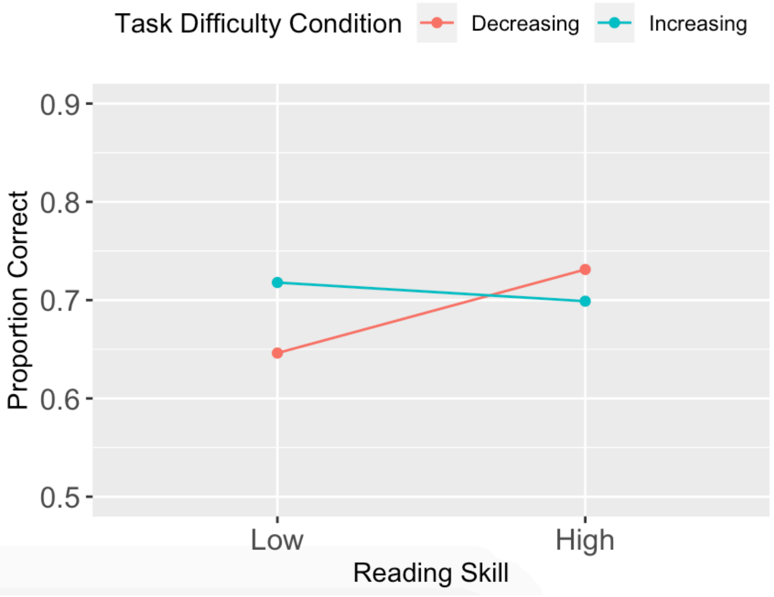


Figure 10. Memory for surface structure in two task difficulty order conditions.

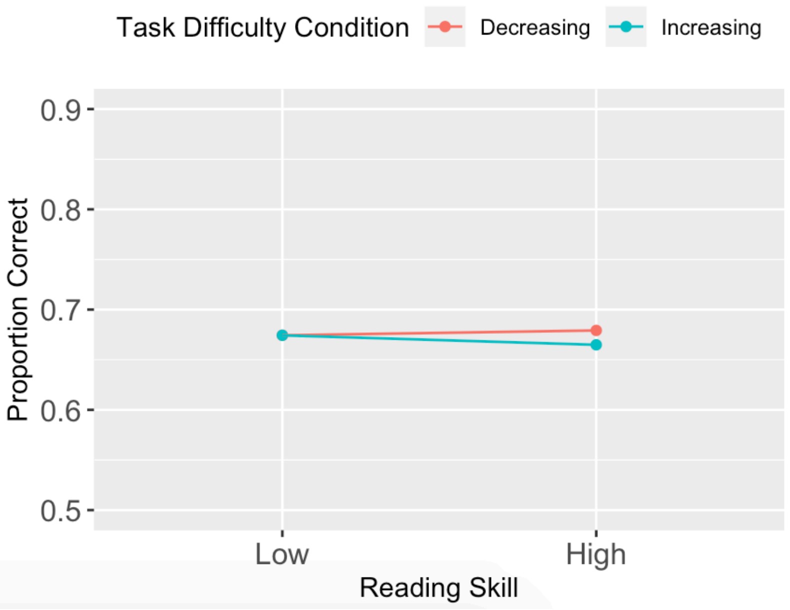


Figure 11. Memory for meaning in two task difficulty order conditions.

Memory for Easy Texts and Difficult Texts. The questions in the recognition test were further separated by the texts they were associated with. More specifically, the questions associated with easy texts were differentiated from the questions associated with difficult texts. Table 6 shows the descriptive statistics (i.e., means and standard deviations) of memory for surface structure and meaning segregating easy and difficult texts in the increasing difficulty order and decreasing difficulty order conditions.

Table 6
Means and standard deviations of high-skill and low-skill readers' memory for surface *structure and meaning associated with different texts*

	Condition	Surface Structure		Meaning	
		Easy	Difficult	Easy	Difficult
High Skill	Decreasing	0.67 (0.28)	0.79 (0.23)	0.63 (0.31)	0.73 (0.25)
	Increasing	0.68 (0.27)	0.71 (0.27)	0.60 (0.32)	0.73 (0.24)
Low Skill	Decreasing	0.61 (0.27)	0.69 (0.28)	0.64 (0.31)	0.71 (0.25)
	Increasing	0.70 (0.23)	0.73 (0.27)	0.64 (0.25)	0.71 (0.27)

Notes. Decreasing = decreasing task difficulty order condition, Increasing = increasing task difficulty order condition; Easy = easy texts, Difficult = difficult texts.

Two mixed-effects linear regressions were performed to examine whether the memory for surface structure associated with easy texts or difficult texts are different in the two task difficulty order conditions. In both models, the independent variables were task difficulty order condition, reading skill level and their interaction. The dependent variable for the first model was the memory for surface structure associated with easy texts. The dependent variable for the second model was the surface structure memory associated with difficult texts. Subjects were specified as a random-effect factor in both models to adjust for the subject variance. Regarding the memory for surface structure associated with easy texts, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (t (156) = 1.49, p = 0.139), the main effect of task difficulty order condition (t (156) = 1.90, p = 0.059), and the main effect of reading skill level (t (310) = 1.61, p = 0.109). Regarding the memory for surface structure associated with difficult texts, the results were statistically nonsignificant for the main effect of task difficulty order condition (t (156) = 1.77, p = 0.079), and reading skill level (t (310) = 2.46, p = 0.015). The interaction between task difficulty order condition and reading skill level was statistically significant (t (156) = 2.13, p = 0.035). Follow-up planned comparisons were performed using repeated measure ANOVAs to compare the memory for surface structure associated with difficulty texts in two task difficulty order conditions for both high-skill readers and low-skill readers. Results indicated high-skill readers' memory for surface structure associated with difficulty texts between two conditions was statistically significant, F (1, 92) = 4.44, p = 0.038. High-skill readers had better memory for surface structure in the decreasing difficulty order condition compared to increasing difficulty order condition. However, low-skill readers' memory for surface structure associated with difficulty texts was not statistically significant, F (1, 64) = 1.05, p = 0.309. The surface structure memory associated with easy and difficult texts in two task difficulty order conditions is shown in Figure 12.

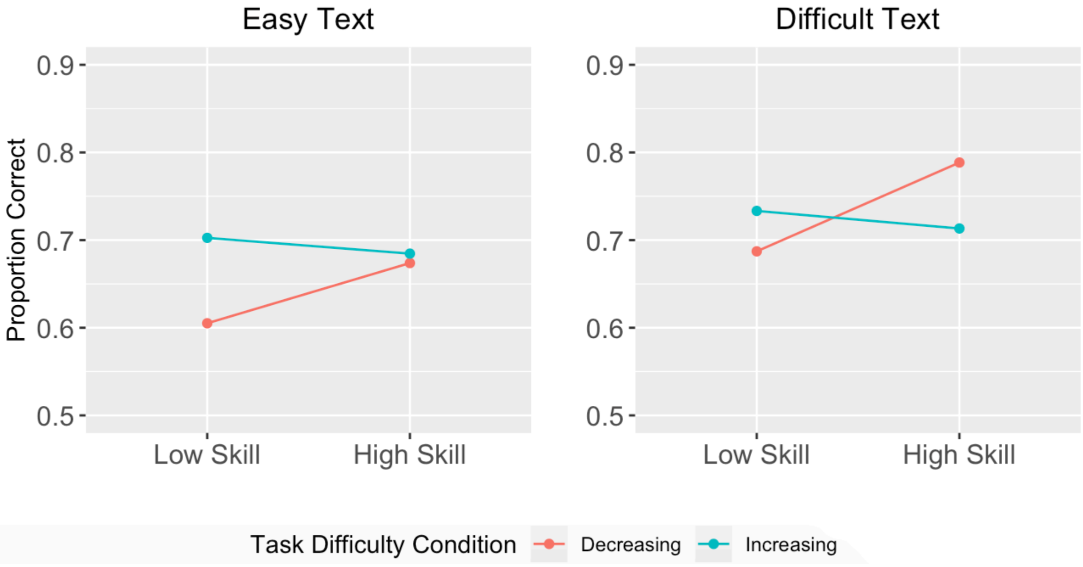


Figure 12. Memory for surface structure associated with easy and difficult texts in two task conditions.

Another two mixed-effects models were performed to analyze the effect of task difficulty order condition, reading skill level and their interaction on the memory for meaning. Regarding the memory for meaning associated with easy texts, the results were not statistically significant for the interaction between task condition and reading skill level *(t* (312) = 0.47, *p* = 0.641), the main effect of task difficulty order condition (*t* (312) = 0.28, *p* = 0.779), and the main effect of reading skill level (*t* (312) = 0.10, *p* = 0.917). Regarding the memory for meaning associated with difficult texts, the results were not statistically significant for the interaction between task difficulty order condition and reading skill (*t* (312) = 0.06, *p* = 0.950), the main effect of task difficulty order (*t* (312) = -0.04, *p* = 0.970), and the main effect of reading skill (*t* (312) = 0.37, *p* = 0.715). The high-skill and low-skill readers' memory for meaning associated with easy and difficult texts in two task difficulty order conditions is shown in Figure 13.

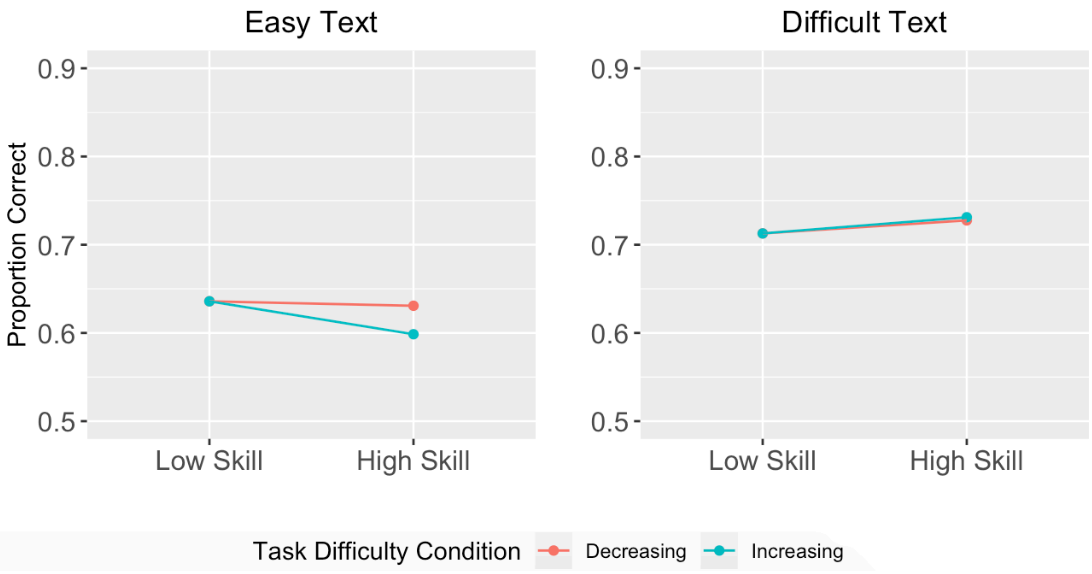


Figure 13. Memory for meaning associated with easy and difficult texts in two task conditions.

Memory for Different Types of Information. In addition to differentiating questions in recognition test according to the type of texts they were related to, the questions were also differentiated according to what type of information they asked about. The questions in the recognition test were divided into three types: questions asking about main idea, questions asking about supporting details and questions asking about information queried by computer agents. Students' memory for surface structure and meaning in two task difficulty order conditions were compared for each type of questions. The descriptive statistics of high-skill and low-skill reader's memory for structure and meaning segregating the three types of questions are shown in Table 7.

Two mixed-effects linear regressions were performed to predict the memory for surface structure and meaning for questions asking about main ideas. The predictors were task difficulty order condition, reading skill level and their interaction. Subjects were specified as a random-effect factor to adjust for the subject variance. Regarding the memory for surface structure, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*t* (312) = 0.36, *p* = 0.772, the main effect of task difficulty order (*t* (312) = 0.15, *p* = 0.884), and the main effect of reading skill level (*t* (312) = 0.16, *p* = 0.147). Regarding the memory for meaning, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*t* (312) = 0.24, *p* = 0.808), the main effect of task difficulty order condition (*t* (312) = 0.19, *p* = 0.847), and the main effect of reading skill level (*t* (312) = 0.14, *p* = 0.892).

Table 7
Means and standard deviations of memory for surface structure and meaning associated with three types of questions

	Condition	Surface Structure			Meaning		
		Main	Support	Queried	Main	Support	Queried
High Skill	Decreasing	0.84 (0.30)	0.74 (0.29)	0.69 (0.34)	0.61 (0.44)	0.71 (0.28)	0.65 (0.35)
	Increasing	0.87 (0.26)	0.72 (0.29)	0.56 (0.36)	0.63 (0.42)	0.67 (0.29)	0.66 (0.32)
Low Skill	Decreasing	0.79 (0.32)	0.64 (0.28)	0.58 (0.38)	0.65 (0.43)	0.67 (0.28)	0.66 (0.33)
	Increasing	0.80 (0.34)	0.72 (0.28)	0.67 (0.33)	0.64 (0.43)	0.69 (0.27)	0.67 (0.27)

Notes. Main = main idea question; Support = supporting detail question, Queried = question asking about information queried by computer agents.

Similarly, two mixed-effects linear regressions were performed to predict memory for surface structure and meaning for questions asking about supporting details. Regarding the memory for surface structure, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*t* (156) = 1.50, *p* = 0.136), the main effect of task difficulty order condition (*t* (156) = 1.45, *p* = 0.149), and the main effect of reading skill level (*t* (308) = 0.25, *p* = 0.804). Regarding the memory for meaning, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*t* (312) = 0.75, *p* = 0.455), the main effect of task difficulty order condition (*t* (156) = 0.55, *p* = 0.586), and the main effect of reading skill level (*t* (311) = 0.37, *p* = 0.711).

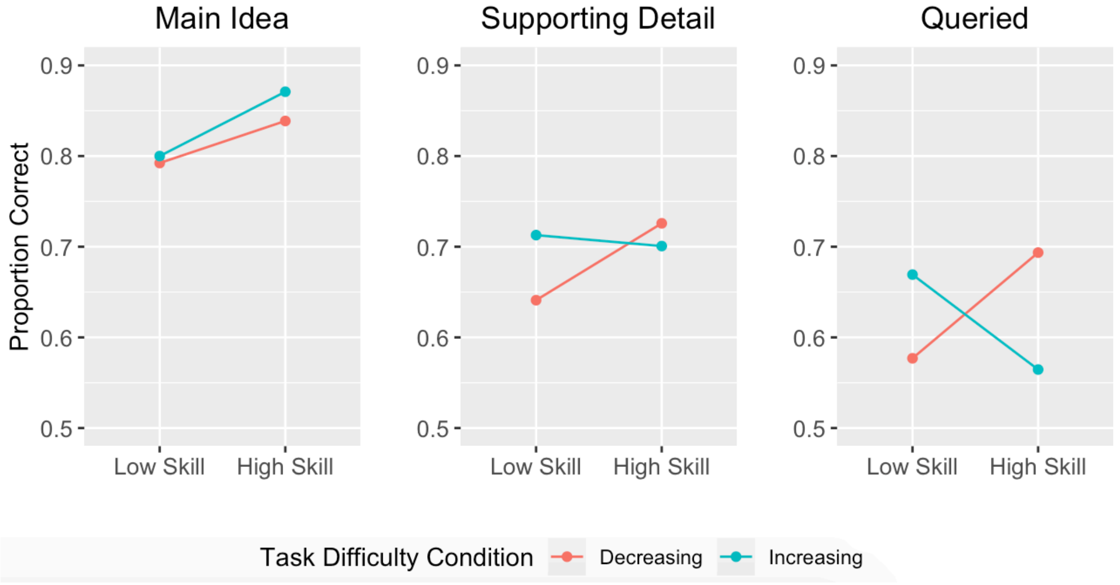


Figure 14. Memory for surface structure associated with three types of questions in two task conditions.

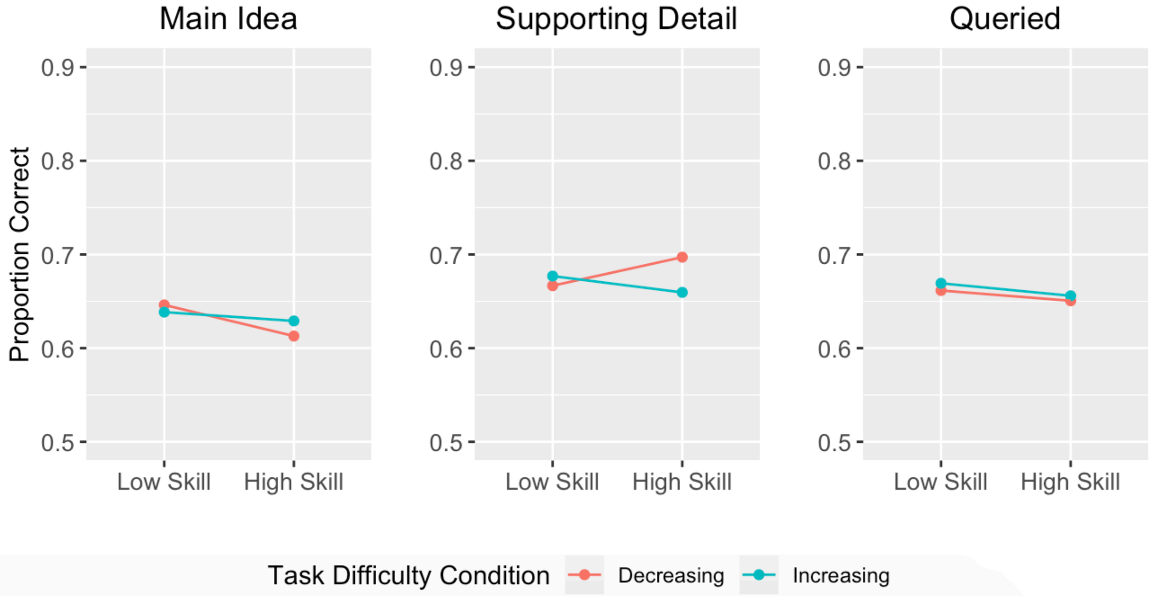


Figure 15. Memory for meaning associated with three types of questions in two task conditions.

Finally, two mixed-effects linear regressions were performed to predict memory for surface structure and meaning for questions asking about information queried by computer agents in AutoTutor. Regarding the memory for surface structure, the results were statistically nonsignificant for the main effect of task difficulty order condition (*t* (312) = 1.50, *p* = 0.135), and the main effect of reading skill level (*t* (312) = 1.85, *p* = 0.065). The interaction between task difficulty order condition and reading skill level was statistically significant (*t* (312) = 2.76, *p* = 0.006). Follow-up planned comparisons were carried out using repeated measure ANOVAs to compare the memory for surface structure regarding queried questions in two task difficulty order conditions for both high-skill readers and low-skill readers. Results indicated the high-skill readers' memory for surface structure between two conditions was statistically significant, *F* (1,184) = 6.44, *p* = 0.012. High-skill readers had better memory for surface structure on querid questions in decreasing difficulty order condition compared to increasing difficulty order condition. Low-skill readers' memory for surface structure between two conditions was not statistically significant, *F* (1,128) = 2.18, *p* = 0.142. Regarding the memory for meaning, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*t* (312) = 0.03, *p* = 0.975), the main effect of task difficulty order condition (*t* (312) = 0.14, *p* = 0.892), and the main effect of reading skill level (*t* (312) = 0.26, *p* = 0.799). The high-skill and low-skill readers' memory for surface structure and meaning on three types of questions are shown in Figure 14 and15.

Effect of Task Difficulty Order on Engagement and Performance during Learning. Engagement was operationalized in two ways in this study. The first operational definition of engagement was time on task. Time on task included text reading time and response time to answer the questions during interactions with AutoTutor. Reading time was at the text level, which means there was one reading time associated with each text for each individual student. Response time was at the question level, with one response time for each question answered by a student. The second operational definition for engagement was the proportion of behaviors within the boundaries (i.e., fast time and slow time) of the engagement zone (Greenberg et al., 2018). The boundaries of a students' engagement zone were computed according to the distribution of response times. Response time was found to be markedly positively skewed, but was normalized with a base-2 logarithmic transformation, which is a common transformation when analyzing response times. Therefore, time boundaries for each text were computed based on the log-transformed distribution of response times for texts and questions associated with the texts. The time boundaries were differentiated between texts because the four texts were at different difficulty levels and the response time of the four texts varied due to different text difficulty. Table 8 shows the descriptive statistics of the log-transformed response times associated with the four texts.

Table 8
Descriptive statistics of log-transformed response time associated with four texts

	M	SD	Skewness	Kurtosis
Text 1	3.46	0.58	1.53	6.84
Text 2	4.16	0.82	0.22	-0.72
Text 3	4.75	0.58	0.76	0.83
Text 4	5.02	0.67	0.43	0.25

In addition to text difficulty, reading skill could also affect response times in answering questions associated with the texts. Therefore, the boundaries of engagement zone were computed text by text by for each individual student. For instance, suppose a student attempted ten questions associated with a text. The lower upper boundaries of the engagement zone for this student on this text would be one standard deviation below and one standard deviation above the mean response time of the ten questions. Each student attempted the questions associated with four texts, thus four lower and four upper boundaries were computed. After the boundaries of the engagement zones were computed, each response time was compared to the relevant boundaries to determine whether the response time was within or beyond the engagement zone. If the response time of an attempt (i.e., answering a question) was within the engagement zone, the attempt was coded as an engaged behavior; otherwise the response time was coded as a disengaged behavior. The descriptive statistics of engagement measures (i.e., reading time, response time and proportion of engaged behaviors) are shown in Table 9.

Table 9
Means and standard deviations performance measures during AutoTutor in two task conditions

Condition	Reading Time	Response Time	Engaged Behavior	Proportion Correct
Decreasing	95.24 (74.79)	25.73 (19.60)	0.68 (0.47)	0.84 (0.36)
Increasing	93.62(64.65)	24.93 (16.37)	0.68 (0.47)	0.82 (0.38)

Notes. Decreasing = decreasing task difficulty order condition, Increasing = increasing task difficulty order condition.

Two mixed-effects linear regressions were performed to predict time on task (i.e., text reading time and answer response time) as a function of task difficulty order condition. Subjects were specified as a random factor to adjust for the subject variance in all the models. The results indicated that the effect of task difficulty order condition on reading time was not statistically significant, *t* (464) = 0.30, *p* = .762. The effect of task difficulty order condition on response time was not statistically significant, *t* (5581) = 1.63, *p* = .103, either. A mixed-effects logistic regression was conducted to analyze the effect of task difficulty order condition on the likelihood of an engaged behavior. It was revealed that the effect of task difficulty order condition on the possibility a behavior being an engaged one was not statistically significant, *z* = -0.12, *p* = .908.

Student's performance accuracy of an attempt was assessed as being correct or incorrect in AutoTutor. Therefore, a mixed-effects logistic regression was conducted to compare students' performance accuracy in two task difficulty order conditions. The mixed-effects logistic regression model predicted the possibility an answer being correct with task difficulty order condition. Subjects were specified as a random factor to adjust for the subject variance. The effect of task difficulty order condition on the likelihood of an answer being correct was not statistically significant, $z = 1.71$, $p = .087$.

The students were further differentiated based on their reading skills, and the interaction effect between task difficulty order and reading skill level on engagement was examined. Table 10 shows the descriptive statistics (i.e., means and standard deviations) of reading time, response time and proportion of engaged behaviors for high-skill readers and low-skill readers.

Table 10
Means and standard deviations of engagement and performance accuracy *for high-skill and low-skill readers in two task conditions*

Skill level	Condition	Reading Time	Response Time	Engagement	Proportion Correct
High Skill	Decreasing	89.67 (53.19)	24.10 (18.97)	0.68 (0.47)	0.86 (0.35)
	Increasing	87.51 (53.69)	23.98 (15.60)	0.68 (0.47)	0.84 (0.36)
Low skill	Decreasing	103.12 (97.55)	28.01 (20.24)	0.69 (0.46)	0.82 (0.39)
	Increasing	102.30 (77.01)	26.27 (17.34)	0.68 (0.47)	0.80 (0.40)

Two mixed-effects linear regressions were performed to predict reading time and response time as a function of task difficulty order condition (i.e., increasing task difficulty and decreasing task difficulty), reading skill level (i.e., high reading skill and low reading skill), and their interaction. In both models, subjects were specified as a random factor to adjust for the subject variance. Regarding reading time, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (t (463) = 0.12, p = 0.901), the main effect of task difficulty order condition (t (462) = 0.03, p = 0.976), and the main effect of reading skill level (t (399) = 1.77, p = 0.076). Regarding response time, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*t* (5581) = 1.59, *p* = 0.111). The main effect of task difficulty order condition was statistically significant (*t* (5581) = 2.00, *p* = 0.045). The response time in decreasing difficulty order condition was longer than that of increasing difficulty order condition ($\beta_{decreasing} = 0.08$). The main effect of reading skill level was also statistically significant (*t* (220) = 2.12, *p* = 0.035). High-skill readers spent shorter time answering questions compared to low-skill readers ($\beta_{high\text{-}skill} = -0.06$).

A mixed-effects logistic regression was conducted to predict the likelihood of an engaged behavior as a function of task difficulty order condition, reading skill level and their interaction. The results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*z* = -0.50, *p* = 0.620), the main effect of task difficulty order condition (*z* = 0.44, *p* = 0.661), and the main effect of reading skill level (*z* = 0.04, *p* = 0.972).

Another mixed-effects logistic regression was performed to predict performance accuracy (i.e., whether an answer is correct or incorrect) as a function of task difficulty order condition, reading skill level and their interaction. The results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*z* = 0.17, *p* = 0.869), and the main effect of task difficulty order condition (*z* = 0.38, *p* = 0.705). The main effect of reading skill level was statistically significant (*z* = 0.78, *p* = 0.005). The students with higher reading skills were more likely to answer a question correctly ($\beta_{high\text{-}skill}$ = 0.31).

The times and performance accuracy related to different types of texts (i.e., easy texts and hard texts) in each task difficulty order condition were further examined. The descriptive statistics are shown in Table 11.

Table 11
Means and standard deviations of engagement and performance accuracy *on different texts in two task conditions for high-skill and low-skill readers*

	Text Type	High-skill Reader		Low-skill Reader	
		Decrease	Increase	Decrease	Increase
Reading Time	Easy	76.62 (53.60)	81.27 (36.90)	90.16 (110.08)	97.59 (50.70)
	Diff	102.57 (49.79)	93.74 (66.01)	116.08 (82.00)	107.01 (96.64)
Response Time	Easy	19.47 (11.33)	20.47 (10.32)	22.73 (11.27)	21.51 (10.83)
	Diff	27.71 (11.88)	26.40 (10.48)	31.92 (15.64)	30.15 (10.84)
Engagement	Easy	0.68 (0.09)	0.69 (0.10)	0.69 (0.08)	0.69 (0.10)
	Diff	0.67 (0.10)	0.67 (0.11)	0.69 (0.08)	0.68 (0.08)
Proportion Correct	Easy	0.92 (0.11)	0.90 (0.13)	0.85 (0.16)	0.86 (0.16)
	Diff	0.81 (0.14)	0.80 (0.14)	0.78 (0.16)	0.74 (0.17)

Notes. Diff = difficult text, Easy = easy text, Decrease = decreasing task difficulty, Increase = increasing task difficulty

Four mixed-effects linear regressions were performed to predict reading time and response time for difficult texts and easy texts with task difficulty order condition (i.e., increasing task difficulty and decreasing task difficulty), reading skill level (i.e., high reading skill and low reading skill), and their interaction. In the models, subjects were specified as a random factor to adjust for the subject variance. Regarding the reading time on easy texts, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*t* (305) = 0.18, *p* = 0.854), the main effect of task difficulty order condition (*t* (305) = 0.41, *p* = 0.684), and the main effect of reading skill level (*t* (305) = 1.53, *p* = 0.127). Regarding the reading time on difficult texts, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*t* (306) = 0.01, *p* = 0.989), the main effect of task difficulty order condition (*t* (306) = 0.33, *p* = 0.739), and the main effect of reading level (*t* (306) = 1.16, *p* = 0.265). Regarding the response time associated with easy texts, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*t* (2790) = 1.92, *p* = 0.055), the main effect of task difficulty order condition (*t* (2788) = 1.77, *p* = 0.077), and the main effect of reading skill level (*t* (331) = 1.05, *p* = 0.293). Regarding the response time associated with difficult texts, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*t* (2645) = 0.43, *p* = 0.661), and the main effect of task difficulty order condition (*t* (2645) = 1.13, *p* = 0.260). The main effect of reading skill level was statistically significant (*t* (227) = 2.43, *p* = 0.016). High-skill readers spent less time answering questions associated with difficult texts compared to low-skill readers ($\beta_{high\text{-}skill}$ = -0.10).

Two mixed-effects logistic regressions were conducted to analyze the effect of task difficulty order condition, reading skill level and their interaction on the likelihood an attempt which was associated with easy tasks or difficult tasks was an engaged behavior. Regarding the engaged behavior associated with easy tasks, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*z* = -0.49, *p* = 0.626), the main effect of task difficulty order condition (*z* = 0.31, *p* = 0.755), and the main effect of reading skill level (*z* = 0.21, *p* = 0.836). Regarding the engaged behavior associated with difficult tasks, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level (*z* = -0.21, *p* = 0.832), the main effect of task difficulty order condition (*z* = 0.31, *p* = 0.756), and the main effect of reading skill level (*z* = -0.16, *p* = 0.875).

Two mixed-effects logistic regressions were performed to analyze the effect of task difficulty order condition, reading skill level and their interaction on the probability of a question which was associated with easy texts or difficult texts was correctly answered. Regarding the answer correctness associated with easy texts, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level ($z = 1.41$, $p = 0.159$), the main effect of task difficulty order condition ($z = -1.06$, $p = 0.287$), and the main effect of reading skill level ($z = 1.90$, $p = 0.057$). Regarding the answer correctness associated with difficult texts, the results were statistically nonsignificant for the interaction between task difficulty order condition and reading skill level ($z = -0.90$, $p = 0.367$), and the main effect of task difficulty order condition ($z = 1.34$, $p = 0.179$). The main effect of reading skill level was statistically significant ($z = 2.32$, $p = 0.020$). High-skill readers were more likely to correctly answer a question associated with difficult texts compared to low-skill readers ($\beta_{high\text{-}skill} = 0.37$).

Moderation Effect of Engagement

In this section, mixed-effects models were conducted to examine the moderation effect of engagement on memory and performance during learning.

Two mixed-effects linear regression models were performed to predict the memory for surface structure and memory for meaning, respectively. In the first model, the memory for surface structure was predicted as a function of task difficulty order condition, reading skill, engagement and their interactions. In the second model, the memory for meaning was predicted as a function of task difficulty order condition, reading skill, engagement and their interactions. In both models, the subjects were specified as a random factor to control for the variance of subjects. The two models were repeated three times by using one of the three engagement measures (i.e., response time, reading time, proportion of engaged behaviors) at a time.

Regarding the memory for surface structure, the three-way interaction between task difficulty order condition, reading skill level and response time were revealed to be statistically nonsignificant (*t* (282) = 0.36, *p* = .720). Given the nonsignificant three-way interaction, the mixed-effects model was modified by removing the three-way interaction term. To be specific, the memory for surface structure was predicted as a function of task difficulty order condition, reading skill level, response time, the two-way interaction between task difficulty order condition and reading skill level, the two-way interaction between task difficulty order and response time, and the two-way interaction between reading skill level and response time. The results were statistically nonsignificant for the two-way interaction between task difficulty order condition and response time (*t* (282) = 0.88, *p* = .308), the two-way interaction between response time and reading skill level (*t* (241) = 1.02, *p* = .309), the main effect of reading skill level (*t* (275) = 0.59, *p* = .556), and the main effect of response time (*t* (289) = 0.38, *p* = .707). The two-way interaction between task difficulty order condition and reading skill level was statistically significant (*t* (154) = 2.70, *p* = .008). The high-skill readers had better memory for surface structure compared to low skill readers in decreasing difficulty order condition ($\beta_{decrease\ *\ high\text{-}skill}$ = 0.26). The main effect of task difficulty order condition was also statistically significant (*t* (303) = 13.63, *p* < .001). Student's memory for surface structure was worse in decreasing difficulty order condition compared to increasing difficulty order condition ($\beta_{decrease}$ = -0.33).

Regarding the memory for meaning, the three-way interaction between task difficulty order condition, reading skill level and response time was not statistically significant (*t* (275) = 1.15, *p* = .250). Given the nonsignificant three-way interaction, the three-way interaction term was removed from the mixed-effects model and the modified model was performed. The results were statistically nonsignificant for the two-way interaction between task difficulty order condition and response time (*t* (277) = 0.39, *p* = .699), the two-way interaction between task difficulty order condition and reading skill level (*t* (155) = 0.32, *p* = .751), the two-way interaction between response time and reading skill (*t* (247) = 1.86, *p* = .064), the main effect of task difficulty order (*t* (303) = 0.26, *p* = .798), the main effect of reading skill level (*t* (277) = 1.52, *p* = .129), and the main effect of response time (*t* (292) = 0.66, *p* = .510).

The previous two mixed-effects linear models were repeated using reading time as the measure of engagement. Specifically, the memory for surface structure and meaning were predicted as a function of task difficulty order condition, reading skill, reading time and the interactions between the three factors. Regarding the memory for surface structure, the results were statistically nonsignificant for the two-way interaction between task difficulty order condition and reading time (*t* (299) = 1.33, *p* = .186), the two-way interaction between task difficulty order condition and reading skill level (*t* (292) = 0.82, *p* = .414), the main effect of task difficulty order condition (*t* (283) = 0.27, *p* = .785), the main effect of reading skill level (*t* (301) = 1.68, *p* = .094), and the main effect of reading time (*t* (298) = 1.63, *p* = .104). The three-way interaction between task difficulty order condition, reading skill level and reading time was found to be statistically significant (*t* (299) = 2.47, *p* = .014). The two-way interaction between reading time and reading skill level was also statistically significant (*t* (298) = 2.16, *p* = .032). Regarding the memory for meaning, the three-way interaction between task difficulty order condition, reading skill level and reading time was indicated to be statistically nonsignificant (*t* (299) = 0.90, *p* = .368). Given the nonsignificant three-way interaction, the three-way interaction term was removed from the mixed-effects model and the modified model was performed. The results were statistically nonsignificant for the two-way interaction between task difficulty order condition and reading time (*t* (300) = 0.86, *p* = .389), the two-way interaction between task difficulty order condition and reading skill level (*t* (150) = 0.34, *p* = .732), the two-way interaction between reading time and reading skill level (*t* (299) = 1.67, *p* = .096), the main effect of task difficulty order condition (*t* (266) = 0.50, *p* = .616), the main effect of reading skill level (*t* (298) = 0.99, *p* = .323), and the main effect of reading time (*t* (299) = 0.48, *p* = .635).

To better understand the statistically significant three-way interaction between task difficulty order condition, reading skill and reading time on the memory for surface structure, the sample used in the model was split into two samples based on reading skill level. After splitting, one sample only included high-skill readers and the other only included low-skill readers. A mixed-effects linear regression was preformed to predict the proportion of recognized surface structure with task difficulty order condition and reading time. The subjects were specified as a random factor to control for their variance. The regression was repeated twice using the two samples, separately. For high-skill readers, there was a statistically significant interaction effect between task difficulty order condition and reading time on the memory for surface structure (*t* (176) = 2.07, *p* = .040). The longer time highs-skill readers spent in the decreasing task difficulty condition, the better memory they had for surface structure ($\beta_{decrease\ *\ reading\ time}$ = 0.26). The main effect of task difficulty order condition was not statistically significant (*t* (175) = 1.37, *p* = .174), nor was the main effect of reading time (*t* (176) = 1.54, *p* = .125). For low-skill readers, the interaction effect between task difficulty order condition and reading time on the memory for surface structure was not statistically significant (*t* (123) = 1.36, *p* = .177). The main effect of task difficulty order condition on the memory for surface structure was not statistically significant (*t* (116) = 0.28, *p* = .782), nor was the main effect of reading time (*t* (122) = -0.14, *p* = .887). The interaction between task difficulty order condition and reading time on memory for surface structure associated with high-skill readers and low-skill readers is shown in Figure 16.

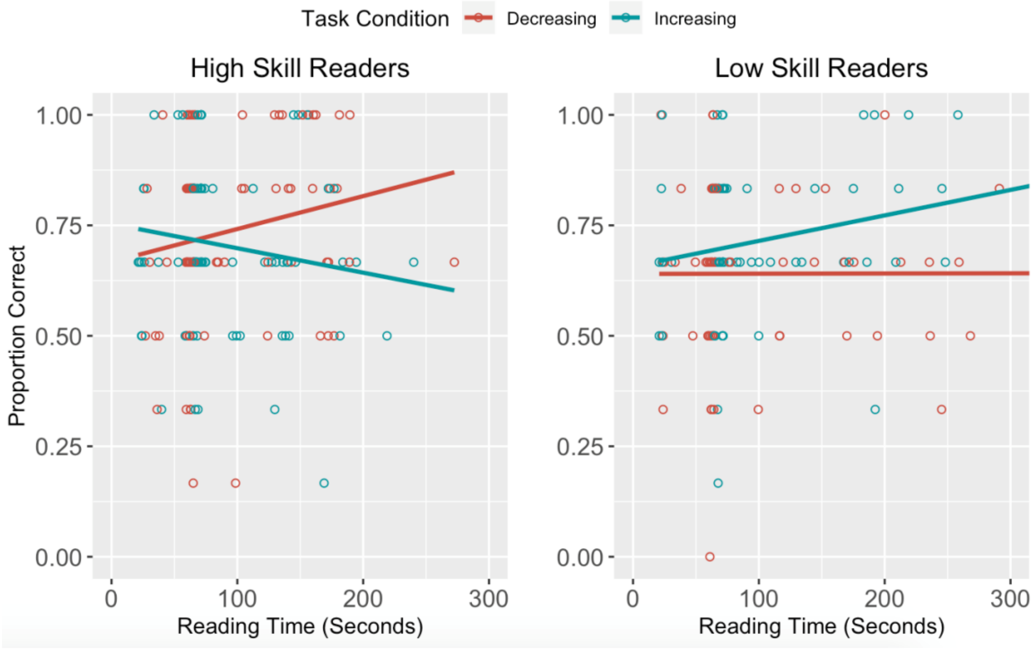


Figure 16. The effect of reading time and task condition on memory for surface structure for high-skill versus low-skill readers.

Two mixed-effects models were conducted to predict memory for surface structure and meaning as a function of task difficulty order condition, reading skill level, the proportion of engaged behavior and their interactions. Regarding the memory for surface structure, the results indicated the three-way interaction between task difficulty order condition, reading skill level and the proportion of engaged behavior was not statistically significant (*t* (294) = 0.56, *p* = .578). Given the nonsignificant three-way interaction, the three-way interaction term was removed from the mixed-effects model and the modified model was performed. The results were statistically nonsignificant for the two-way interaction between task difficulty order condition and proportion of engaged behavior (*t* (299) = 0.72, *p* = .475), the two-way interaction between proportion of engaged behavior and reading skill level (*t* (298) = 1.22, *p* = .223), the main effect of task difficulty order condition (*t* (299) = 1.04, *p* = .301), the main effect of reading skill level (*t* (299) = 1.26, *p* = .210), and the main effect of proportion of engaged behavior (*t* (302) = 0.91, *p*

= .362). The two-way interaction between task difficulty order condition and reading skill level was statistically significant (*t* (152) = 2.65, *p* = .009). High-skill readers had better memory for surface structure compared to low skill readers in decreasing difficulty condition ($\beta_{decrease\ *\ high\text{-}skill}$ = 0.26).

Regarding the memory for meaning, the results indicated the three-way interaction between task difficulty order condition, reading skill level and the proportion of engaged behavior was not statistically significant (*t* (298) = 1.63, *p* = .104). Given the nonsignificant three-way interaction, the three-way interaction term was removed from the mixed-effects model and the modified model was performed. The results were statistically nonsignificant for the two-way interaction between task difficulty order condition and proportion of engaged behavior (*t* (302) = 0.90, *p* = .368), the two-way interaction between task difficulty order condition and reading skill level (*t* (153) = 0.24, *p* = .813), the two-way interaction between proportion of engaged behavior and reading skill level (*t* (153) = 0.24, *p* = .813), the main effect of task difficulty order condition (*t* (301) = 0.86, *p* = .392), the main effect of reading skill level (*t* (302) = 0.05, *p* = .963), and the main effect of proportion of engaged behavior (*t* (303) = 0.58, *p* = .561).

To analyze the moderation effect of engagement on performance accuracy in AutoTutor, a mixed-effects linear regression was performed in which the proportion of correct answers was predicted by task difficulty order condition, reading skill level, engagement and their interactions. The subjects were specified as a random factor to control for the variance of subjects. The same model was run three times using three different engagement measures (i.e., response time, reading time, proportion of engaged behaviors). With response time as the measure for engagement, the three-way interaction between task difficulty order condition, reading skill level and response time was not statistically significant (*t* (285) = 1.08, *p* = .280). Given the nonsignificant three-way interaction, the three-way interaction term was removed from the mixed-effects model and the modified model was performed. The results were statistically nonsignificant for the two-way interaction between task difficulty order condition and response time (*t* (287) = 1.17, *p* = .242), the two-way interaction between task difficulty order condition and reading skill level (*t* (144) = 0.32, *p* = .752), the two-way interaction between response time and reading skill level (*t* (225) = 0.97, *p* = .333), the main effect of task difficulty order condition (*t* (302) = 0.33, *p* = .743), and the main effect of reading skill level (*t* (270) = 1.81, *p* = .072). The main effect of response time on performance accuracy was statistically significant (*t* (281) = 5.23, *p* < .001). The students who spent longer time answering questions answered fewer questions correctly in AutoTutor ($\beta_{Response\ time}$ = -0.50).

The second model used reading time as the measure for engagement. This model predicted the proportion of correct answers as a function of task difficulty order condition, reading skill level, reading time and their interactions. The results indicated the three-way interaction between task condition, reading skill and reading time was not statistically significant (*t* (302) = 0.27, *p* = .788). Given the nonsignificant three-way interaction, the three-way interaction term was removed from the mixed-effects model and the modified model was performed. The results were statistically nonsignificant for the two-way interaction between task difficulty order condition and reading time (*t* (303) = 1.01, *p* = .315), the two-way interaction between task difficulty order condition and reading skill level (*t* (303) = 0.18, *p* = .859), the main effect of task difficulty order condition (*t* (303) = 1.33, *p* = .186), and the main effect of reading time (*t* (303) = 0.05, *p* = .958). The main effect of reading skill level on performance accuracy was statistically significant (*t* (303) = 3.81, *p* < .001). High-skill readers answered more questions correctly compared to low-skill readers ($\beta_{high\text{-}skill}$ = 0.44). The two-way interaction between reading time and reading skill level on performance accuracy was also statistically significant (*t* (303) = 3.14, *p* = .002). Compared to low-skill readers, high-skill readers who spent longer time reading the texts answered a lower proportion of questions correctly ($\beta_{high\text{-}skill\ *\ reading\ time}$ = -0.33).

The third model used the proportion of engaged behavior as the measure for engagement. This model predicted the proportion of correct answers as a function of task difficulty order condition, reading skill level, proportion of engaged behavior and their interactions. Results indicated a statistically nonsignificant three-way interaction between task difficulty order condition, reading skill level and proportion of engaged behavior (*t* (302) = 1.13, *p* = .258). Given the nonsignificant three-way interaction, the three-way interaction term was removed from the mixed-effects model and the modified model was performed. The results were statistically nonsignificant for the two-way interaction between task difficulty order condition and proportion of engaged behavior (*t* (303) = 0.52, *p* = .603), the two-way interaction between task difficulty order condition and reading skill level (*t* (303) = 0.04, *p* = .972), the two-way interaction between proportion of engaged behavior and reading skill level (*t* (303) = 1.03, *p* = .260), the main effect of task difficulty order condition (*t* (303) = .40, *p* = .693), the main effect of reading skill level (*t* (303) = 0.81, *p* = .422), and the main effect of proportion of engaged behavior (*t* (303) = 0.08, *p* = .936).

Chapter 5 Discussion

In this dissertation, an experiment was conducted to investigate the impact of task difficulty order on memory, engagement, and performance during learning in a reading comprehension intervention with computer agents. The learning tasks were organized in two contrasting orders: an increasing task difficulty order and a decreasing task difficulty order. Easy learning tasks were followed by difficult tasks with the increasing difficulty order. Difficult learning tasks were presented before easy tasks with the decreasing difficulty order. The learning tasks were presented in a conversation-based intelligent tutoring system called AutoTutor, which utilizes three-way tutorial conversations (i.e., trialogues) between a teacher agent, a peer agent and a human student. Each student interacted with AutoTutor on two lessons. One lesson was in decreasing difficulty order and the other was in increasing difficulty order. Students' learning behaviors were tracked and recorded by AutoTutor. The students also took a reading comprehension test before the AutoTutor session and a recognition test after the AutoTutor session. The reading comprehension test assessed students' reading skills whereas the recognition test evaluated their memory of the texts presented in AutoTutor.

This dissertation had three major goals. The first goal was to examine the impact of task difficulty order on memory, engagement, and performance during learning. The second goal was to investigate the interaction between task difficulty order and individual difference, specifically reading skill, in terms of their impact on memory, engagement, and performance during learning. The third goal was to explore the moderation effect of engagement on memory and performance during learning.

Goal 1: Does task difficulty order influence memory, engagement, and performance during learning?

Comparisons were made in students' memory for surface structure and meaning in the two task difficulty order conditions. It was found that there was no significant overall difference between increasing difficulty order condition and decreasing difficulty order condition on students' memory for surface structure. Similarly, the difference between two task conditions was nonsignificant regarding students' memory for meaning. The questions in the recognition tests were further differentiated into three categories: questions asking about main ideas, questions asking about supporting ideas, and questions asking about the information queried by the computer agents. Students' memory for surface structure and meaning for each type of question in the two task conditions were compared. The results indicated that students' memory for surface structure was not significantly different between the two task conditions for any type of questions. There were also no differences in memory for meaning.

There was an interesting finding in a direct comparison in the memory for surface structure versus the meaning. Earlier studies reported memory for surface structure is quickly lost, whereas memory for meaning is longer lasting (Begg, 1971; Sachs, 1974). Similarly, surface structure is retained in short-term memory but decays quickly and is much less available in long-term memory, whereas meaning is preserved in long-term memory (Kintsch, 1970; Baddeley, 1966). This would suggest that there should be better memory for meaning than surface structure in a delayed memory condition. The results of the present study did not support this prediction. In the decreasing difficulty order condition, the average proportion of recognized surface structure was 0.70 while the average proportion of recognized meaning was 0.68. This difference was not statistically significant in repeated-measure one-way ANOVA, F (1, 157) =

0.94, p = 0.335. In the increasing difficulty order condition, the average proportion of recognized surface structure and meaning was 0.71 and 0.67, respectively, which was a significant difference (F (1, 157) = 4.05, p = 0.046), but in the opposite direction from the prediction. Graesser and Mandler (1975) reported long-term memory for surface structure depended on how the information was processed at the input. Specifically, there would be poor memory for surface structure if the subjects processed primarily the more semantic, conceptual information. However, if the subjects paid attention to the wording and phrasing of the discourse, the recognition for surface structure could be considerably above chance (see also Long, 1994). It appears that the students in this study did pay close attention to the texts while interacting with AutoTutor, presumably because the focus of AutoTutor was on comprehension training. That is, the main goal of the lessons was to improve comprehension of texts so it would be prudent for students to concentrate on the wording of the material. If the focus had been on learning science or other informational topics then there may have been poorer memory for surface structure. It is also possible that the trialogues between computer agents and human effectively directed the students' attention to the wording and phrasing of the texts.

In addition to memory, the dissertation also compared the engagement and performance accuracy during learning between two task difficulty order conditions. There were two operational definitions for engagement in this dissertation: time on task and behavior within the zone of engagement (Greenberg et al., 2018). Time on task was measured by reading time (i.e., time spent on reading a text) and response time (i.e., time spent on answering a question). The proportion of behaviors within zone of engagement was computed based on response time. Specifically, boundaries of the engagement zone were created according to the distribution of a student's response time on questions related to a text. An attempt was counted as a behavior within zone of engagement if the response time of the attempt was within the boundaries. The results of mixed-effects models indicated no statistically significant differences between the two task conditions in terms of reading time, response time, or proportion of engaged behaviors. The performance accuracy in the two task difficulty order conditions were also compared, and again the results indicated no statistically significant difference between the two task conditions.

Goal 2: Does reading skill interact with task difficulty order in analyses of memory, engagement and performance during learning?

Analyses were performed to investigate how and to what extent students' reading skills interacted with task difficulty order in terms of their effect on memory, engagement and performance during learning. The technique used to differentiate students' reading skills was a cluster analysis. We examined the four measures generated by the maze subtest (Sabatini et al., 2019) and selected two of them that measured performance accuracy and reading speed. The cluster analysis generated two student clusters. One cluster read faster and performed more accurate, and it was labeled as high-skill readers. The other cluster, low-skill readers, read slower and with less accuracy. The results of mixed-effects modelling indicated a significant interaction between the task difficulty order condition and reading skill with respect to the effect on the memory for surface structure. Low-skill readers had better memory for surface structure in the increasing task difficulty order condition, whereas high-skill readers' memory for surface structure was not significantly different between the two conditions. No significant interaction was found between task difficulty order condition and reading skill with respect to the effect on the memory for meaning.

The finding regarding the memory for surface structure for low-skill readers was consistent with our predictions. The students had the best surface structure memory when they were first provided with easy texts (and tasks) that were within their zone of proximal development. By the time low-skill readers reached the difficult tasks, they had already been exposed to similar tasks that focused on the same reading comprehension strategies. The difficult tasks were challenging but apparently still within their zone of proximal development due to the prior learning. However, the difficult tasks were probably too difficult and beyond low-skill readers' zone of proximal development when they were provided at the very beginning of the learning process. Low-skill readers might not be able to understand the difficult texts in this condition, so their memory was significantly worse than when they were in increasing difficulty order condition.

In addition to the overall interaction effect found between task difficulty order condition and reading skill level, the texts were separately examined by difficult level. Students' memory for difficult texts, as well as their memory for easy texts, was compared between two task difficulty order conditions. A significant interaction effect was found between task difficulty order condition and reading skill level on the memory for surface structure associated with difficult texts. High-skill readers had better memory for surface structure associated with difficult texts in the decreasing difficulty order condition, whereas low-skill readers' memory for surface structure associated with difficult texts were not significantly different between two conditions. The interaction between task difficulty order condition and reading skill was not significant for easy texts. However, low-skill readers had better memory for surface structure associated with easy texts in the increasing difficulty order condition. The easy texts were probably within the ZPD of low-skill readers, which meant they could understand the texts well and even memorize surface structure details that were presented at the beginning of the learning tasks. When the difficult texts were presented after the easy materials, low-skill readers may have gained confidence from the prior relevant learning experience. On the other hand, the difficult learning materials might be too challenging for the low-skill readers when they were presented at the beginning. In this situation, the low-skill readers did not have a good understanding of the texts which resulted in poor memory of the information in the texts. Additionally, the difficult tasks, when presented first, may have made these students anxious, given the absence of confidence-building easy tasks. This may also be why low-skill readers had better memory in the decreasing difficulty order condition than increasing difficulty order condition. For high-skill readers, their memory for difficult texts was better in decreasing difficulty order condition than increasing difficulty order condition. Perhaps the easy text and associated tasks at the beginning failed to stimulate the high-skill readers and this lower engagement dampened performance in subsequent tasks.

High-skill and low-skill readers' memory for different types of information was also examined. Specifically, the questions in recognition test were categorized into three types: questions about main idea, questions about supporting details, and questions about the information queried by computer agents. Regarding the memory for surface structure, there was a significant interaction between the task difficulty order condition, reading skill, and question category (see Figure 14). Low-skill readers' memory for surface structure associated with queried questions were not significantly different between two conditions, whereas high-skill readers were found to recognize more surface structure associated with queried questions in the decreasing difficulty order condition. This interaction effect is consistent with the interaction effect for difficult texts. After further examining the difference between queried questions supporting detail questions, it turned out that the queried questions also asked about supporting details. Given this, the difference between supporting detail questions and queried questions is whether the details were emphasized by computer agents or not. This finding suggests that the memory for surface structure might be associated with the amount of attention directed to the details. The three types of questions are getting more detail oriented from main ideas question, to supporting details question, to queried question. Meanwhile, the strength of the interaction between reading skill and task condition for the three types of questions is increasing, as is shown in Figure 14.

There was no significant interaction between reading skill and task difficulty order condition on engagement. The main effect of task condition on engagement was also nonsignificant. The only factor that was found to affect engagement was reading skill. High-skill readers spent less time answering questions compared to low-skill readers, particularly on questions associated with difficult texts. Given the significant interaction effect between reading skill and task condition found on memory, a similar pattern might be expected for engagement. However, the engagement was not found to be affected by the interaction of task condition and reading skill, which is somewhat counterintuitive.

Perhaps the validity of the engagement measures was inadequate in current study. Previous research has categorized engagement into three types: behavioral engagement, emotional engagement and cognitive engagement (Fredricks et al., 2004). Behavioral engagement refers to students' participation and involvement in learning tasks, including effort and attention (Fredricks et al., 2004; Fredricks & McColskey, 2012). Emotional engagement refers to the affective reactions such as boredom, happiness, and anxiety (Fredricks et al., 2004; Skinner & Belmont, 1993). Cognitive engagement refers to psychological investment in learning tasks such as how students manage and control effort towards mastering knowledge and skills taught in school (Lamborn, Newmann, & Wehlage, 1992; Pintrich & De Groot, 1990). However, Pekrun & Linnenbrink-Garcia (2012) suggested that there is some conceptual overlap between cognitive and behavioral engagement. Very little consensus has been reached regarding how to operationalize and measure each aspect of engagement, so multiple measures are often recommended (Shernoff, 2013). Engagement in this study falls into the category of behavioral engagement and was operationalized with two measures. One measure is time on task, including text reading time and question response time. Spanjers, Burns and Wagner (2008) studied the relationship between time on task and engagement self-reported by students. They found the correlation between time on task and self-reported engagement was weak. This might be the reason the time on task measures were not found to be affected by task condition and reading skill. Another measure used for engagement was behaviors within the zone of engagement (Greenberg et al., 2018). The boundaries of engagement zone were computed based on the distribution of response time from each student. The times beyond the boundaries were regarded as out of the zone of engagement. However, it is not clear where the actual boundaries of an engagement zone should be. The way the zone of engagement was computed was exploratory. It is possible that there was an inaccurate overlap between the zone of engagement created in this study and the actual zone of engagement.

The comprehension test of learning from AutoTutor showed a nonsignificant interaction between task difficulty order condition and reading skill. The main effect of task difficulty order was also nonsignificant. However, high-skill readers answered a higher portion of questions correctly than low-skill students, as expected. Therefore, the effect of task difficulty order on performance accuracy in AutoTutor does not align with the effect of task difficulty order condition on memory. Most of the questions in AutoTutor are deep questions that help students read texts at deeper levels of comprehension (Graesser et al., 2012; Graesser, Ozuru, & Sullins, 2010), and require reasoning to fully answer. Fuzzy tracing theory suggests that reasoning can be independent of memory (Brainerd & Reyna, 1992; Reyna & Brainerd, 1995). Perhaps this is why the effect of task difficulty order on performance accuracy in AutoTutor is not consistent with the effect of task difficulty order on memory.

It is worth noting that the investigation of time on task and performance accuracy for easy texts and difficult texts separately validated the task difficulty measured used in this study. As is shown in Table 11 on page 61, the reading time and response time related to difficult texts were longer than that of easy texts for both high-skill readers and low-skill readers in two task difficulty conditions. Similarly, the performance accuracy associated with easy texts was higher than the accuracy associated with difficult texts in two task difficulty conditions.

Goal 3: Does engagement moderate effects on memory and performance during learning?

The moderation effect of engagement on memory was not statistically significant using response time or behaviors within zone of engagement as indicators of engagement. However, the moderation effect of engagement on memory was significant when reading time was used as the indicator of engagement, as can be seen in Figure 16 on page 67. High-skill readers recognized more surface structure when they spent more time reading in the decreasing difficulty order condition. Conversely, high-skill readers had better memory for surface structure when the reading time was longer in the increasing difficulty order condition. Longer reading time was associated with better memory for surface structure in the decreasing difficulty order condition and instead of increasing difficulty order condition. High skill-readers may have been bored with the initial easy task, so their longer reading times may be an indicator for mind-wandering and disengagement. Low-skill readers, on the other hand, had better memory for surface structure as they spent more time reading the texts in the increasing difficulty order condition. In the decreasing difficulty order condition, the trending line is flat. It is possible that the difficult texts were too challenging for low-skill readers so they could not understand them despite spending more time reading. In the increasing difficulty order condition, low-skill readers were first provided with the easier text and questions on the same topic. With prior learning, low-skill readers were better prepared for the difficult texts and they might also be less anxious. Given this, low-skill readers could benefit from longer reading and memorize more surface structure in the increasing difficulty order condition.

In summary, students' reading skill level was found to interact with task difficult order regarding their impact on memory. It was also found that reading time, as an indicator of engagement, moderated the effect of task difficulty order on students' memory. Specifically, students with lower reading skills had better memory for the texts in the increasing difficulty order condition compared to the decreasing difficulty order condition. This finding supports the theory of zone of proximal development (ZPD) (Vygotsky, 1978). The students with lower reading skill were kept within their ZPD by starting with easy learning materials and then moving to more difficult learning materials. On the other hand, when low-skill students were given difficult learning materials initially, they did not learn as well, which suggests the difficult learning materials were outside of low-skill students' ZPD. The students with higher reading skill were found to have better memory for difficult texts in the decreasing difficulty order condition compared to the increasing difficulty order condition. The difference between the two task conditions was probably due to different engagement states. The easy learning materials might be too easy for high-skill students and led to boredom, and there may have been some residual boredom when the students began the difficult learning materials after the easy learning materials. The negative consequences of boredom may have affected high-skill readers longer in the increasing difficulty condition, where easy materials were provided at the beginning, compared to the decreasing difficulty order condition where easy materials were given after difficulty materials.

Limitations and Future Directions

One limitation of this study is the validity of the engagement measures. Engagement in this study refers to behavioral engagement (i.e., students' participation and involvement in learning tasks) (Fredricks et al., 2004; Fredricks & McColskey, 2012). It was operationalized with time on task (i.e., reading time, response time), and behaviors within personal zone of engagement. Although reading time was found to moderate the effect of task difficulty order condition on memory, the validity of reading time as the measure of engagement remains unclear. Future research that include self-report surveys of engagement will help validate the behavioral engagement measures. This dissertation investigated the domain of reading comprehension and found an interesting interaction effect between task difficulty order and reading skill on memory. To see if these results generalize, future research can use the same approach but in a different domain.

Conclusion

The findings of this dissertation offer some empirical support for the Zone of Proximal Development in the domain of reading comprehension. It further supports the idea that ITSs interventions should consider students' individual differences and emphasizes that ITSs should provide learning materials at the right level of difficulty for each individual. Simply put, students are not able to learn if the learning materials are too difficult and beyond their ZPD. On the other hand, students may get distracted and disengaged if the learning materials are too easy. Adaptive learning systems should start with assessments that gauge students' knowledge states and skill level, and then provide learning materials and instructions accordingly. For instance, ITSs can start a reading comprehension intervention by providing some short texts and learning tasks with mixed levels of difficulty to diagnose students' reading skills. The system can then feed learning materials to students based on the diagnosis. The learning materials and learning tasks can be used for both instruction and assessment. Students' performance on prior learning tasks provide useful information to ITSs, which can be used to determine what learning materials should be provided for the following instruction.

References

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist51*(4), 355.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The journal of the learning sciences, 4*(2), 167-207.

Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *The Quarterly Journal of Experimental Psychology, 18*, 362–365.

Balaban, N. (1995). Seeing the child, knowing the person. *To become a teacher*, 49-57.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19*(2), 185-224.

Beal, C. R., Arroyo, I., Cohen, P. R., Woolf, B. P., & Beal, C. R. (2010). Evaluation of AnimalWatch: An intelligent tutoring system for arithmetic and fractions. *Journal of Interactive Online Learning, 9*(1), 64-77.

Beal, C. R., Walles, R., Arroyo, I., & Woolf, B. P. (2007). Online tutoring for math achievement testing: A controlled evaluation. *Journal of Interactive Online Learning, 6*(1), 43-55.

Begg, I. (1971). Recognition memory for sentence meaning and wording. *Journal of Verbal Learning and Verbal Behavior, 10*(2), 176-181.

Belanich, J., Sibley, D., & Orvis, K. L. (2004). *Instructional characteristics and motivation features of a PC-based game (ARI Research Report 1822)*. U.S. Army Research Institute for the Behavioral and Social Sciences: Alexandria, VA.

Berk, L., & Winsler, A. (1995). Vygotsky: His life and works" and" Vygotsky's approach to development. *Scaffolding children's learning: Vygotsky and early childhood learning*, 25-34. National Association for the Education of Young Children: Washington, DC

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16.

Brainerd, C. J., & Reyna, V. F. (1992). The memory independence effect: What do the data show? What do the theories claim? *Developmental Review, 12*(2), 164-186.

Britton, B. K., & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of educational Psychology, 83*(3), 329-345.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). Package 'clvalid': Validation of clustering results. *Journal of Statistical Software, 25*, 1-22.

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). Effectiveness of reading and mathematics software products: findings from two student cohorts. NCEE 2009-4041. *National Center for Education Evaluation and Regional Assistance*.

Chambers, B., Slavin, R. E., Madden, N. A., Abrami, P., Logan, M. K., & Gifford, R. (2011). Small-group, computer-assisted tutoring to improve reading outcomes for struggling first and second graders. *The Elementary School Journal*, *111*(4), 625–640.

Chambers, B., Slavin, R. E., Madden, N. A., Abrami, P. C., Tucker, B. J., Cheung, A., & Gifford, R. (2008). Technology infusion in success for all: reading outcomes for first graders. *The Elementary School Journal*, *109*(1), 1–15.

Cheung, A. C. & Slavin, R. E. (2013). Effects of educational technology applications on reading outcomes for struggling readers: A best-evidence synthesis. *Reading Research Quarterly*, *48*(3), 277–299.

Chi, M. T. H. (1996). Constructing Self-Explanations and Scaffolded Explanations in Tutoring. *Applied Cognitive Psychology, 10*(7), 33–49.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist, 49*(4), 219-243.

Cohen, P. A., Kulik, J. A., & Kulik, C. L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*, 237–248.

Craig, S. D., Hu, X., Graesser, A. C., Bargagliotti A. E., Sterbinsky, A., Cheney, K. R., & Okwumabua, T. (2013). The impact of a technology-based mathematics after-school program using ALEKS on student's knowledge and behaviors. *Computers & Education, 68*, 495-504.

Cronbach, L. (1957). The Two Disciplines of Scientific Psychology. *American Psychologist*, *12*(11), 671-684.

Csikszentmihalyi, M. (1988). The flow experience and its significance for human psychology. In M. Csikszentmihalyi & I. S. Csikszentmihalyi (Eds.), *Optimal experience: Psychological studies of flow in consciousness* (pp. 15-35). New York, NY: Cambridge University Press.

Czikszentmihalyi, M. (1990). Flow: The psychology of optimal experience. New York, NY: Harper & Row.

Dance, K. A., & Neufeld, R. W. (1988). Aptitude-treatment interaction research in the clinical setting: A review of attempts to dispel the" patient uniformity" myth. *Psychological Bulletin, 104* (2), 192-213.

D’Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145-157.

Doignon, J. P., & Falmagne, J. C. (1999). Knowledge spaces. New York: Springer-Verlag.

Doignon, J. P., & Falmagne, J. C. (2016). Knowledge spaces and learning spaces. *New Handbook of Mathematical Psychology, 2*, 274-321.

Durlach, P. J., & Spain, R. D. (2014). *Framework for Instructional Technology: Methods of Implementing Adaptive Training and Education* (No. ARI-RR-1335). Fort Belvoir, VA: Army Research Institute for Behavioral and Social Sciences.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., ... & Emery, D. (2007). Effectiveness of reading and mathematics software products: Findings from the first student cohort. Lyon, France: HAL

Falmagne, J. C., Albert, D., Doble, C., Eppstein, D., & Hu, X. (Eds.). (2013). *Knowledge spaces: Applications in education*. Berlin, Germany: Springer Science & Business Media.

Feng, S., D’Mello, S., & Graesser, A. C. (2013). Mind wandering while reading easy and difficult texts. *Psychonomic bulletin & review, 20*(3), 586-592.

Feuerstein, R., Falik, L., Rand, Y., & Feurerstein, R. S. (2002). The dynamic assessment of cognitive modifiability. *The learning propensity assessment device: Theory, instruments and techniques*. Jerusalem, Israel: ICELP Press.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research, 74*(1), 59-109.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S.

Christenson, A. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 763–782). New York: Springer.

Graesser, A. C. (2009). Cognitive scientists prefer theories and testable principles with teeth. *Educational Psychologist, 44*(3), 193-197

Graesser, A.C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education, 26*,124-132.

Graesser, A.C., Cai, Z., Baer, W.O., Olney, A.M., Hu, X., Reed, M., & Greenberg, D. (2016). Reading comprehension lessons in AutoTutor for the Center for the Study of Adult Literacy. In S.A. Crossley and D.S. McNamara (Eds.). *Adaptive educational technologies for literacy instruction* (pp. 288-293). New York: Taylor & Francis Routledge.

Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, & H. L. Swanson (Eds.), *APA educational psychology handbook, Vol. 3. Application to learning and teaching* (pp. 451-473). Washington, DC, US: American Psychological Association.

Graesser, A. C., Forsyth, C. M., & Lehman, B. A. (2017). Two heads may be better than one: learning from computer agents in conversational trialogues. *Teachers College Record, 119*(3), 1-20.

Graesser, A. C., Hu, X., & McNamara, D. S. (2005). Computerized learning environments that incorporate research in discourse psychology, cognitive science, and computational linguistics. *Experimental Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer.* Washington, DC: American Psychological Association.

Graesser, A. C., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science, 23*(5), 374-380.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 180-192.

Graesser, A. C., & Mandler, G. (1975). Recognition memory for the meaning and surface structure of sentences. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(3), 238-248.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in cognitive science, 3*(2), 371-398.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115*(2), 210-229.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational researcher, 40*(5), 223-234.

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z. & Hu, X. (2007). Using LSA in AutoTutor: Learning through mixed initiative dialogue in natural language. In T. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 243–262). Mahwah, NJ: Erlbaum.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, *9*(6), 495–522.

Graesser, A., Ozuru, Y., Sullins, J.: What is a good question? In: McKeown, M.G., Kucan, L. (eds.) *Threads of Coherence in Research on the Development of Reading Ability*, (pp. 112–141). Guilford, New York (2009b)

Graesser, A.C., Rus, V. & Hu, X. (2017). Instruction based on tutoring. In R.E. Mayer & P.A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 460-482). New York: Routledge Press.

Greenberg, D., Graesser, A.C., Frijters, J.C., Lippert, A.M., & Talwar, A. (2018). *Using AutoTutor to track performance and engagement in a reading comprehension intervention for adult literacy students.* Manuscript submitted for publication.

Halpern, D. F., Millis, K., Graesser, A. C., Butler, H., Forsyth, C., & Cai, Z. (2012). Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. *Thinking Skills and Creativity, 7*(2), 93-100.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100-108.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651-666.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR), 31*(3), 264-323.

Ji, X. R., Beerwinkle, A., Wijekumar, K., Lei, P., Malatesha Joshi, R., & Zhang, S. (2018). Using latent transition analysis to identify effects of an intelligent tutoring system on reading comprehension of seventh-grade students. *Reading and Writing*, *31*, 1-19

Keislar, E. R., & Stern, C. (1970). Differentiated instruction in problem solving for children of different mental ability levels. *Journal of Educational Psychology*, *61*, 445-450.

Kintsch, W. (1970). *Learning, memory, and conceptual process*. New York: Wiley.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction integration model. *Psychological review, 95*(2), 163-182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, UK: Cambridge university press.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly, 10*, 62–102.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science, 36*(5), 757-798.

Landauer, T. K., & Psotka, J. (2000). Simulating text understanding for educational applications with Latent Semantic Analysis: Introduction to LSA. *Interactive Learning Environments, 8*(2), 73-86.

Lamborn, S., Newmann, F., & Wehlage, G. (1992). The significance and sources of student engagement. *Student Engagement and Achievement in American Secondary Schools*, 11-39.

Linderholm, T., Everson, M. G., Van Den Broek, P., Mischinski, M., Crittenden, A., & Samuels, J. (2000). Effects of causal text revisions on more-and less-skilled readers' comprehension of easy and difficult texts. *Cognition and Instruction, 18*(4), 525-556.

Long, D. L. (1994). The effects of pragmatics and discourse style on recognition memory for sentences. *Discourse Processes, 17*(2), 213-234.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901-918.

Malone, T. W., Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, Learning and Instruction* (pp. 223-253). Hillsdale, NJ: Lawrence Erlbaum Associates.

Massimini, F., & Carli, M. (1988). The systematic assessment of flow in daily experience. In M. Csikszentmihalyi & I. S. Csikszentmihalyi (Eds.), *Optimal experience: Psychological studies of flow in consciousness* (pp. 266-287). New York, NY: Cambridge University Press.

McCarthy, K., Likens, A., Kopp, K., Perret, C., Watanabe, M., & McNamara, D. (2018). The "LO"-down on grit: non-cognitive trait assessments fail to predict learning gains in iSTART and W-Pal. In *LAK 2018, Workshop on Online Learning & Non-Cognitive Assessment at Scale,* Sydney, AU.

McDaniel, M. A., Einstein, G. O., Dunay, P. K., & Cobb, R. E. (1986). Encoding difficulty and memory: Toward a unifying theory. *Journal of Memory and Language, 25*(6), 645-656.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix.* Cambridge, M.A.: Cambridge University Press.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1-43.

McNamara, D. S., O'Reilly, T. P., Best, R. M., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research, 34*(2), 147-171.

McShane, S., & Travaglione, A. (2007). *Organisational behaviour on the Pacific Rim*. North Ryde, NSW: McGraw-Hill Australia.

Meyer, B. J. F., & Lei, P. (2017). Web-based text structure strategy instruction improves seventh graders' content area reading comprehension. *Journal of Educational Psychology*, *109*(6), 741–760.

Meyer, B. J., Wijekumar, K., Middlemiss, W., Higley, K., Lei, P. W., Meier, C., & Spielvogel, J. (2010). Web-based tutoring of the structure strategy with or without elaborated feedback or choice for fifth-and seventh-grade readers. Reading Research Quarterly, 45(1), 62-92.

Mills, C., Graesser, A., Risko, E. F., & D'Mello, S. K. (2017). Cognitive coupling during reading. *Journal of Experimental Psychology: General, 146* (6), 872-907.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education, 24*(4), 427-469.

O'Brien, E. J., & Myers, J. L. (1985). When comprehension difficulty improves memory for text. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(1), 12.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse processes, 43*(2), 121-152.

O'Reilly, T. P., Sinclair, G. P., & McNamara, D. S. (2004). iSTART: A web-based reading strategy intervention that improves students' science comprehension. In Kinshuk, D. G. Sampson, & P. Isaías (Eds.), *Proceedings of the IADIS International Conference Cognition and Exploratory Learning in Digital Age: CELDA 2004* (pp. 173–180). Lisbon, Portugal: IADIS.

Orvis, K. A., Horn, D. B., & Belanich, J. (2008). The roles of task difficulty and prior videogame experience on performance and motivation in instructional videogames. *Computers in Human behavior, 24*(5), 2415-2433.

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19*(3), 228-242.

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds*.), Handbook of research on student engagement* (pp. 259–282). New York: Springer.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33.

Reschly, A., & Christenson, S. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. Christenson, A. Reschly, & C.Wylie (Eds.), *Handbook of research on student engagement* (pp. 3-19). Berlin: Springer.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and individual Differences, 7*(1), 1-75.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*(2), 249-255.

Sabatini, J., Weeks, J., O' Reilly, T., Bruce, K., Steinberg, J., & Chao, S.-F. (2019). *SARA Reading Components Tests, RISE forms: Technical adequacy and test design, 3rd edition* (Research Report No. RR-19-36). Princeton, NJ: Educational Testing Service. https://doi.org10.1002/ets2.12269

Sachs, J. S. (1974). Memory in reading and listening to discourse. *Memory & Cognition, 2*(1), 95-100.

Schaffer, H. R. (2006). *Key concepts in developmental psychology.* Los Angeles, CA: Sage.

Schell, J. (2014). *The Art of Game Design: A book of lenses.* Boca Raton, FL: CRC Press.

Shernoff, D. J. (2013). Measuring student engagement in high school classrooms and what we have learned. In *Optimal learning environments to promote student engagement* (pp. 77-96). Springer, New York, NY.

Skanes, G. R., Sullivan, A. M., Rowe, E. J. & Shannon, E., (1974), Intelligence and transfer: Aptitude by treatment interactions. *Journal of Educational Psychology, 66*, 563-568.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571.

Smallwood, J., McSpadden, M., & Schooler, J. W. (2007). The lights are on but no one's home: Meta-awareness and the decoupling of attention when the mind wanders. *Psychonomic Bulletin & Review, 14*(3), 527-533.

Smith, J. E. (2001). *The effect of the Carnegie Algebra Tutor on student achievement and attitude in introductory high school algebra* (Doctoral dissertation). Virginia Polytechnic Institute and State University, Blacksburg.

Snow, R. (1987). Aptitude complexes. *Aptitude, Learning and Instruction*, *3*, 11-34.

Snow, R. (1989). Cognitive-conative aptitude in learning. In R. Kanfer, P.L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation and methodology: The Minnesota symposium on learning and individual differences* (pp. 435-474). Hillsdale, NJ: Erlbaum.

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of consulting and clinical psychology, 59* (2), 205-216.

Spanjers, D. M., Burns, M. K., & Wagner, A. R. (2008). Systematic direct observation of time on task as a measure of student engagement. *Assessment for Effective Intervention, 33*(2), 120-126.

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology, 105*(4), 970-987.

Stenner, A. J. (2006). *Measuring reading comprehension with the Lexile framework*. Durham, NC: Metametrics, Inc. Retrieved from http://files.eric.ed.gov/fulltext/ED435977.pdf.

Vanlehn, K. (2006). The behavior of tutoring systems. *International journal of artificial intelligence in education, 16*(3), 227-265.

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

Van Velsor, E., & McCauley, C. D. (2004). Our view of leadership development. In C. D. McCauley & E. Van Velsor (Eds.), *The center for creative leadership: Handbook of leadership development* (pp. 1–22). San Francisco, CA: Jossey-Bass.

Voss, J. F., & Silfies, L. N. (1996). Learning from history text: The interaction of knowledge and comprehension skill with text structure. *Cognition and Instruction, 14* (1), 45-68.

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental process.* Cambridge, Mass: Harvard University Press

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236-244.

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*(2), 85-120.

Weaver, C. A., & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition, 23*(1), 12-22.

Wickelgren, W. A. (1972). Coding, retrieval, and dynamics of multitrace associative memory. In Gregg, L.W. (Ed.), *Cognition in learning and memory.* Oxford, England: John Wiley & Sons.

Wijekumar, K. K., Meyer, B. J. F., & Lei, P. (2012). Large-scale randomized controlled trial with 4th graders using intelligent tutoring of the structure strategy to improve nonfiction reading comprehension. *Educational Technology Research and Development: ETR & D*, *60*(6), 987–1013.

Wijekumar, K. K., Meyer, B. J. F., & Lei, P. (2013). High-fidelity implementation of web-based intelligent tutoring system improves fourth and fifth graders content area reading comprehension. *Computers & Education*, *68*, 366–379.

Wijekumar, K. K., Meyer, B. J. F., & Lei, P. (2017). Web-based text structure strategy instruction improves seventh graders' content area reading comprehension. *Journal of Educational Psychology*, *109*(6), 741-760.

Wijekumar, K. K., Meyer, B. J. F., Lei, P., Cheng, W., Ji, X., & Joshi, R. M. (2017). Evidence of an Intelligent Tutoring System as a Mindtool to Promote Strategic Memory of Expository Texts and Comprehension With Children in Grades 4 and 5. *Journal of Educational Computing Research*, *55*(7), 1022–1048.

Winegar, L. T. (1988). Child as cultural apprentice: An alternative perspective for understanding zone of proximal development. *The Genetic Epistemologist, 16*(3), 31-38.

Wolf, M., Barzillai, M., Gottwald, S., Miller, L., Spencer, K., Norton, E., ... & Morris, R.

(2009). The RAVE-O intervention: Connecting neuroscience to the classroom. *Mind, Brain, and Education, 3*(2), 84-93.

APPENDICES

A. Informed Consent Form

Principal Investigator: Arthur Graesser

Study Title: Reading Comprehension

Institution: University of Memphis

The following information is provided to inform you about the research project and your participation in it. Please read this form carefully and feel free to ask any questions you may have about this study and the information given below. You will be given an opportunity to ask questions, and your questions will be answered.

Your participation in this research study is voluntary. You are also free to withdraw from this study at any time. In the event new information becomes available that may affect the risks or benefits associated with this research study or your willingness to participate in it, you will be notified so that you can make an informed decision whether or not to continue your participation in this study.

For additional information about giving consent or your rights as a participant in this study, please feel free to contact the IRB at 901-678-4758 or email irb@memphis.edu.

1. **Purpose of the study:**

You are being asked to interact with a computer program that is designed to help you read better. In order to design this computer as best we can, we need to record how you use the computer so we can make improvements. This will help researchers and educators tailor reading programs to characteristics of the reader.

2. **Description of procedures to be followed and approximate duration of the study:**

When you sign the consent, you are allowing researchers to record and analyze your data taken from the computer. However, the data will not be shown to the public and will remain in the data analyses research team. The ID number that you receive after consenting to this study will match your name to this data. This number will be held by researchers and used only to identify which data can be used for analysis. After this data collection is completed, any information on your identity will be destroyed so that your data will be anonymous.

3. **Expected costs:**

There are no discernible costs or risks to participation in this study unless the participants have a negative affect for performing less than expected. The encouragement of the experimenters should mitigate this concern.

4. **Description of the discomforts, inconveniences, and/or risks that can be reasonably expcted as a result of participation in this study:**

There are no major discomforts beyond reading assignments.

5. **Compensation in case of study-related injury:**

University of Memphis does not have a fund set aside for compensation in the case of study related injury.

6. Anticipated benefits from this study:

a) The potential benefits to science and humankind that may result from this study are that there are expected to be improved methods to help reading. This would also advance theories of learning, cognition, and reading in addition to helping individual participants.

b) The potential benefits to you from this study are credit in an Introductory Psychology class and learning about reading strategies.

7. Alternative treatments available:

There is no alternative available to help participants improve their reading.

8. Compensation for participation: There is no punishment for participation in the study. However, you will receive course credit for your participation if you are part of the participant pool that is organized by the Department of Psychology.

9. Circumstances under which the Principal Investigator may withdraw you from study participation:

There are no circumstances to withdraw you from the study unless there are problems in viewing computer displays or listening to instructions.

10. What happens if you choose to withdraw from study participation:

If you choose to withdraw from the study, you will not be penalized in any way.

11. Contact Information.

If you should have any questions about this research study or possible injury, please feel free to contact Arthur Graesser at 901-678-4857 or 901-678-2742. For questions regarding the research subjects' rights, the Chair of the Institutional Review Board for the Protection of Human Subjects should be contacted at 901-678-2533.

12. Confidentiality.

All efforts, within the limits allowed by law, will be made to keep the personal information in your research record private. However, there may be unexpected extenuating circumstances, such that total privacy cannot be promised. Your information may be shared with U of M or the government, such as the University of Memphis University Institutional Review Board, Federal Government Office for Human Research Protections, if you or someone else is in danger or if we are required to do so by law.

I understand the following facts about participation in this study. All information collected in the study is confidential, within the limits allowed by law, and my name will only be used to identify my data from the recordings of my interacting with the computer. After collecting the data, my name will no longer be needed, and records will be destroyed that link my name with my ID number. No one else will ever see my name associated with the data under any circumstances because these associations will be permanently destroyed. I give Dr. Graesser permission to use information from the collected data for journal publications, conferences, and presentations. A numeric code will be used as identification on data collection materials. Once data are collected, this code will be used for maintenance and analysis of data. Pseudonyms will be used in publications and conference papers. According to guidelines of the American Psychological Association, after 5 years all data will be destroyed.

STATEMENT BY PERSON AGREEING TO PARTICIPATE IN THIS STUDY

I have read this informed consent document and the material contained in it has been explained to me verbally. I understand each part of the document, all my questions have been answered, and I freely and voluntarily choose to participate in this study.

Subject's Name (print): ____________________

Subject's Signature: ________________________ Date: ________________________

Investigator's Signature: _____________________ Date: ________________________

B. IRB Approval

From: <irb@memphis.edu>

Date: Fri, May 11, 2018 at 11:34 AM

Subject: 2255 - Renewal: Approval - Renewal

To: graesser@memphis.edu

Institutional Review Board

Office of Sponsored Programs

University of Memphis

315 Admin Bldg

Memphis, TN 38152-3370

PI: Arthur Graesser

Co-Investigator:

Advisor and/or Co-PI:

Department: Psychology

Study Title: Understanding the cognitive and motivational profiles of struggling adult readers and developing effective and engaging literacy programs to address their literacy learning needs

IRB ID: 2255

Submission Type: Renewal

Level of Review: Expedited

IRB Meeting Date:

Decision: Approved

Approval Date: May 11, 2018

Expiration Date: May 11, 2019

The IRB has reviewed the renewal request.

Approval of this project is given with the following obligations:

1. If this IRB approval has an expiration date, an approved renewal must be in effect to continue the project prior to that date. If approval is not obtained, the human consent form(s) and recruiting material(s) are no longer valid and any research activities involving human subjects must stop.

2. When the project is finished or terminated, a completion form must be completed and sent to the board.

3. No change may be made in the approved protocol without prior board approval, whether the approved protocol was reviewed at the Exempt, Expedited or Full Board level.

4. Exempt approval are considered to have no expiration date and no further review is necessary unless the protocol needs modification.

Thank you,

James P. Whelan, Ph.D.

Institutional Review Board Chair

The University of Memphis.

Note: Review outcomes will be communicated to the email address on file. This email should be considered an official communication from the UM IRB.

C. Demographic Survey

Instructions: Please answer the following questions to the best of your ability.

1. Assigned ID: ____________________
2. What is your current age in years? ___________________
3. What is your gender?
    - o Male
    - o Female
4. What's your Ethnicity (or Race)?
    - o White
    - o Hispanic or Latino
    - o Native American or American Indian
    - o Asian / Pacific Islander
    - o Other, please specify________
5. What is your native language? ________________
6. How many years have you learned English? _________________

D. Recognition Test

1. Which sentence appeared in the lesson text?

   A. It has been on Earth ever since life forms started to grow beyond a single cell.

   B. Ever since life forms started to grow beyond a single cell, it has been on Earth.

   C. It has been on Earth ever since life forms started to divide beyond a single cell.

   D. Ever since life forms started to divide beyond a single cell, it has been on Earth.

2. Which sentence appeared in the lesson text?

   A. Sometimes genes get damaged, and a cell starts acting the wrong way.

   B. A cell starts acting the wrong way when genes get damaged.

   C. Sometimes genes get damaged, and a cell starts dividing wildly.

   D. A cell starts dividing wildly when genes get damaged.

3. Which sentence appeared in the lesson text?

   A. Since then the country has spent more than $100 billion on the fight.

   B. The country has spent more than $100 billion on the fight since then.

   C. Since then the country has spent more than $100 billion on cancer research.

   D. The country has spent more than $100 billion on cancer research since then.

4. Which sentence appeared in the lesson text?

   A. Facebook is the obvious news powerhouse among the social media sites.

   B. The obvious news powerhouse among social media sites is Facebook.

   C. Facebook is the obvious news powerhouse among news pathways.

   D. The obvious news powerhouse among news pathways is Facebook.

5. Which sentence appeared in the lesson text?

   A. Finally, our recent survey revealed that social media does not always facilitate conversation around important issues of the day.

B. Finally, our recent survey revealed that conversation around important issues of the day is not always facilitated by social media.

C. Finally, our recent survey revealed that entertainment news does not always facilitate conversation around important issues of the day.

D. Finally, our recent survey revealed that conversation around important issues of the day is not always facilitated by entertainment news.

6. Which sentence appeared in the lesson text?

A. Pew Research found that in 2014, 14% of social media users posted their own photos of news events to a social networking site.

B. In 2014, Pew Research found that 14% of social media users posted their own photos of news events to a social networking site.

C. Pew Research found that in 2014, 14% of social media users posted their own photos of news events to Facebook.

D. In 2014, Pew Research found that 14% of social media users posted their own photos of news events to Facebook.

7. Which sentence appeared in the lesson text?

A. Schedule multiple alarms and Sleeper Time lets you listen to music for a set time as you fall asleep.

B. Schedule multiple alarms and Sleeper Time lets you listen to music for a set time as you wake up.

C. Sleeper Time lets you listen to music for a set time as you fall asleep and you can schedule multiple alarms.

D. Sleeper Time lets you listen to music for a set time as you wake up and you can schedule multiple alarms.

8. Which sentence appeared in the lesson text?

A. Can download up to 10 free ringtones.

B. Up to 10 free ringtones can be downloaded.

C. Can download up to 5 free ringtones.

D. Up to 5 free ringtones can be downloaded.

9. Which sentence appeared in the lesson text?

A. Displays location of calling number, phone number, and names of callers.

B. Location of calling number, phone number, and names of callers are displayed.

C. Displays location of calling number, phone number, and pictures of callers.

D. Location of calling number, phone number, and pictures of callers are displayed.

10. Which sentence appeared in the lesson text?

A. Walking and running are low-cost, easy-to-do anywhere, year-round activities.

B. Two low cost, easy-to-do anywhere, year-round activities are walking and running.

C. Walking and running are low-cost, easy-to-do anywhere, social activities.

D. Two low-cost, easy-to-do anywhere, social activities are walking and running.

11. Which sentence appeared in the lesson text?

A. In order to get any benefit from a workout, it has to be one that you enjoy and will do day after day.

B. In order to get any benefit from a workout, it has to be one that you will do day after day and you enjoy.

C. In order to get any benefit from a workout, it has to be one that you enjoy and is not repetitive.

D. In order to get any benefit from a workout, it has to be one that is not repetitive, sand you enjoy.

12. Which sentence appeared in the lesson text?

A. Walking is good exercise for those who are just starting to workout, or for those with health problems.

B. For those who are just starting to workout, or for those with health problems, walking is good exercise.

C. Walking is good exercise for those who are just starting to workout, or for those who are overweight.

D. For those who are just starting to workout, or for those who are overweight, walking is good exercise.