Journal of Literacy and Technology
Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

An Empirical Evaluation of Reading Comprehension Tablet Software Utilizing the Question Generation Strategy

Dr. Kevin Browne
Mohawk College
kevinbrowne@gmail.com

Dr. Christopher Anand
McMaster University
anandc@mcmaster.ca

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

Abstract

Motivated by the substantial socioeconomic issues associated with low literacy skills, we developed a tablet application to help adults improve their reading comprehension skills. Though reading comprehension is a complex and multifaceted skill, studies have shown that teaching students metacognitive strategies can help them improve their reading comprehension skills. We created an iPad application which attempts to teach users the question generation strategy. We built another iPad application as a control that only allowed users to practice their reading comprehension skills; this application did not attempt to teach a metacognitive skill. We tested the applications with 48 undergraduate and graduate student participants from McMaster University. The application which aimed to teach the question generation strategy resulted in a statistically significant improvement in reading comprehension performance relative to the control application. We conclude that tablet software can be used to teach the question generation strategy, and propose directions for future work.

Keywords: literacy, reading comprehension, metacognitive strategies, adult literacy, gamification, dynamic difficulty adjustment

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

Introduction

The high prevalence of low literacy skills in adults and associated employment, economic and health impacts has been extensively documented by international and national organizations (Long, 2001; Vernon, 2007; Yin et al., 2006; Barr-Telford, Nault & Pignal, 2005; Kutner, Greenberg, & Baer, 2005). Programs to address low literacy show disappointing enrollment and drop-out rates. The drop-out rate for those participating in literacy organization programming is 30%, and under 10% of Canadians who could benefit from a literacy program enroll (Long, 2001), with those who do not enroll citing employment, financial, childcare and transportation constraints.

Tablet computers have a touch screen interface that may be simpler relative to a keyboard and mouse, could lower costs relative to a human instructor, and allow a user to access interactive educational content remotely (without feeling embarrassment by revealing their low literacy skills to an instructor). As such, these devices appear to provide an opportunity for novel and disruptive approaches to the problem of low adult literacy. Indeed, an Apple Vision video from 1988 envisaged many tablet features and suggested they would be helpful for adult literacy education (McGuinnessPublishing, 2011).

This opportunity led the authors to conduct an exploratory study into the effectiveness of tablet software incorporating gamification and serious game design approaches. Three iPad applications for teaching punctuation and homophone literacy concepts were designed, developed and tested with adult-literacy program clients of the Brant Skills Centre. Groups of participants received instruction of the relevant concepts both with the iPad applications and through more traditional lecture-style instruction. This study was able to show that game design elements could increase learner engagement, and that these tablet software applications are likely best suited for the drill and practice phases of learning.

That the applications were most suited for drill and practice, while not a goal of the previous research, is a natural result of seeking aspects of literacy acquisition which would most likely

demonstrate the effectiveness of gamification and serious game design. It was easiest to design experiments around easily defined subproblems (punctuation, homophones). Although the previous applications were designed in consultation with Brant Skills Centre instructions, and incorporated Brian Cambourne's and Kayne Toukonen's thoughts on learning (Rushton, Eitelgeorge & Zickafoose, 2003; Toukonen, 2011), they were not designed to address the most challenging issues identified by literacy researchers, nor were they designed to teach literacy independently of the instructor.

In this study, we sought to demonstrate that independent learning of a core reading strategy can also be facilitated by tablet software, which really opens up the possibility of remote learning, increasing flexibility and reducing costs of adult education.

During the previous study we noted that traditional lecture-style instruction had a distinct advantage in teaching the concepts due to the dynamism of the instructor. In-the-moment adjustments based on participant reactions, learning tailored to individual needs, and emotional sensitivity combined to make instructors effective. Our applications came closest to this effectiveness when they exhibited dynamic characteristics, for example, a gradually increasing difficulty level, or corrective feedback after an incorrect response to a question.

This work builds on our previous efforts and insights, going deeper by combining learning with practice, teaching more abstract concepts, and exploring the use of dynamic difficulty adjustment in adult-literacy software. Of all the subproblems of literacy teaching, we decided that teaching strategies for reading comprehension best covered these goals.

Given the overlap in terms, and to help better situate our work and goals, we will define reading, literacy, digital literacy and reading comprehension. Reading is a cognitive process of recognizing symbols and constructing meaning and/or understanding (reading comprehension). Traditionally literacy is defined as the ability to read and write (specifically, to read and write text). However, the

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

concept is sometimes expanded to include the ability to understand information, and to use multimedia, numbers, mathematics, and technology. Digital literacy is the ability of a person to produce, evaluate, find and communicate on digital platforms (for example, the Internet). Sometimes digital literacy is included in the broader definitions of literacy, in recognition of the importance of digital platforms to modern life. Closely associated with the idea of digital literacy is a concern about a digital divide, that is the inequity in access to digital platforms and communication technologies.

Reading comprehension is the ability to read, decode and comprehend text. Reading comprehension is a complex, multifaceted and creative process, about which much is known. Reviewing this knowledge is beyond the scope of this paper, but it is important to take individual differences into account. To give a short list, reading comprehension is dependent upon individual differences in: working memory (Daneman & Carpenter, 1980), vocabulary knowledge (Cain, Oakhill, & Lemmon, 2004), background knowledge (Pearson, Hansen & Gordon, 1979), phonology (Bishop & Adams, 1990), interest level (Sousa & Oakhill, 1996), inference-making ability (Cain & Oakhill, 1998), text-anomaly resolution ability (Yuill, Oakhill & Parkin, 1989) and cultural background (Johnson, 1981). In addition to the individual, texts also vary widely in subject and style, from more narrative texts such as novels, to technical texts, such as scientific journal papers, and there is a rich literature to mine for approaches to tablet learning, but for our first tablet application in this area, we need a tight focus.

The majority of reading past the primary grades and the majority of reading required by adults to succeed in life and at work involves expository text (Stanovich & Siegel, 1994). Expository text is intended to explain or describe something. The ability to comprehend expository text will only become more important with society's increasing dependence on technology (Lapp, Flood & Ranck-Buhr, 1995). For these reasons, we focus our work towards the comprehension of expository text. Further, we focus

on the recall of information presented directly in the text itself, i.e., on answering who, what, when, where, why, and how, without requiring inference or interpretation by the reader.

Metacognitive reading strategies are considered key to improving reading comprehension within the literature on the subject (Shang & Chang-Chien, 2010; McNamara, 2012; Jetton & Dole, 2004). One way that proficient readers are different from struggling readers is in their application of metacognitive reading strategies (Baker & Brown, 1984). Such strategies involve the reader reflecting on and consciously thinking about what they have read in various different ways, for example, by attempting to visualize (Bell & Lindamood, 1981) or summarize (Bean & Steenwyk, 1984) a passage of text. Proficient readers will employ these strategies before, after, and during the reading of a passage of text (Paris, Wasik & Turner, 1991). Numerous experimental results have shown that struggling readers can improve their performance if they are taught to apply these metacognitive strategies during learning sessions conducted over a period of time (Shang & Chang-Chien, 2010; Wong & Jones, 1982; Bereiter & Bird, 1985). Some experimental results have shown that struggling readers can improve their reading comprehension in a single learning session (Gambrell & Bales, 1986; Bereiter & Bird, 1985). While reading comprehension strategies have previously been taught using software successfully, for example iSTART (McNamara, Levinstein & Boonthum, 2004), we are unaware of any study documenting using tablet software to teach a reading comprehension strategy.

Though our work is inspired by metacognitive strategies which have been in-use for decades, there is more recent research and curriculum approaches that we should also note to help position our work. Teaching multiple reading comprehension strategies designed to improve inference ability to middle school students has shown to be effective (Barth & Elleman, 2017). Elementary schools have used educational literacy software to improve student engagement and test scores, and to promote growth in

metacognitive abilities of students (Smith, 2016). Using spaced repetition software has been shown to be effective at helping ESL students improve performance on Test of English for International Communication (TOEIC) scores (Bower, 2016). First grade students who were briefly taught a rehearsal strategy aimed at improving verbal working memory performed better than a control group that was not (Peng & Fuchs, 2015). Though our work does not directly build on any of these works specifically, it utilizes similar ideas and approaches in a different way, and we hope in this sense that it fills a gap in the current literature.

The *question generation* reading comprehension strategy involves having the learner generate and answer questions in the process of reading the text (Cohen, 1983; Rosenshine, 1997). Our focus on reading comprehension of who, what, where, when, why and how information in a text lends itself naturally to the question generation strategy.

Dynamic difficulty adjustment is a game design concept that involves modifying the difficulty of a game while it is being played (Hunicke, 2005), in contrast to for example selecting a level of difficulty for the game before play begins. *Gamification* can be defined as the “usage of game design elements to motivate user behavior in non-game contexts" (Deterding, 2011). Dynamic difficulty adjustment fits this definition and can be used to make an experience engaging to a wide spectrum of different users (Missura, 2015).

With these motivations and after having consulted the cited literature, we arrived at the following primary research questions:

- Can reading comprehension performance be improved by teaching the question generation strategy using tablet software?

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

- Can incorporating dynamic difficulty adjustment and gamification design elements in reading comprehension tablet software result in high user engagement?

In this paper, we present the design and experiment results for two iPad applications we developed. One of the iPad applications attempts to teach the user the question generation strategy, and incorporates several gamification design elements, in particular dynamic difficulty adjustment. The other iPad application was created for use by a control group, and allows users to practice their reading comprehension skills without teaching them the question generation strategy and without incorporating dynamical difficult adjustment and most of the gamification elements. The experiment participants were McMaster University students, selected according to a Research Ethics Board approved plan.

We did find a *statistically significant improvement in reading comprehension* over the control group when using the application incorporating metacognitive strategies and dynamic difficulty adjustment. We did not find evidence for an improvement in user engagement in the experiment group over the control group; this may have been due to issues with the design of our application which caused some participant frustration.

This experiment design does not allow independent quantitative effectiveness evaluation of the reading strategy versus gamification, because our conception of responsively "teaching" the reading strategy required additional software elements and we could not see a practical way of implementing responsiveness without some level of gamification, and the most natural approach was dynamic difficulty adjustment. As we will explain, however, the user survey and qualitative comments allows us to make judgements about the merits of the two features.

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

In Section 2, we present the design of the iPad applications. In Section 3, we discuss the design of the usability experiment that took place. In Section 4, we analyze the results of this study, and in Section 5, we discuss our conclusions and provide directions for future work.

Application Design

Two applications were designed and built for the iPad. The *experiment application* was built to improve reading comprehension skills using the question generation strategy, and improve engagement using dynamic difficulty adjustment and gamification. The *control application* was built to be used by a control group, and as a result is only meant to provide a chance for practicing reading comprehension via a series of passages and questions. The control application does not teach the question generation strategy, and does not feature dynamic difficulty adjustment or more sophisticated gamification included in the experiment application.

Experiment Application

The experiment application was designed to teach the question generation strategy to the user, and to facilitate high user engagement, primarily by using dynamic difficulty adjustment.

Design Overview

The first screen that the user accesses is the topic selection screen shown in Figure 1. The topic selection screen allows the user to select a topic to read about. The topic selection screen initially allows the user to select from one of two topics, but the user has the ability to unlock more options as a reward based on their performance in the application.

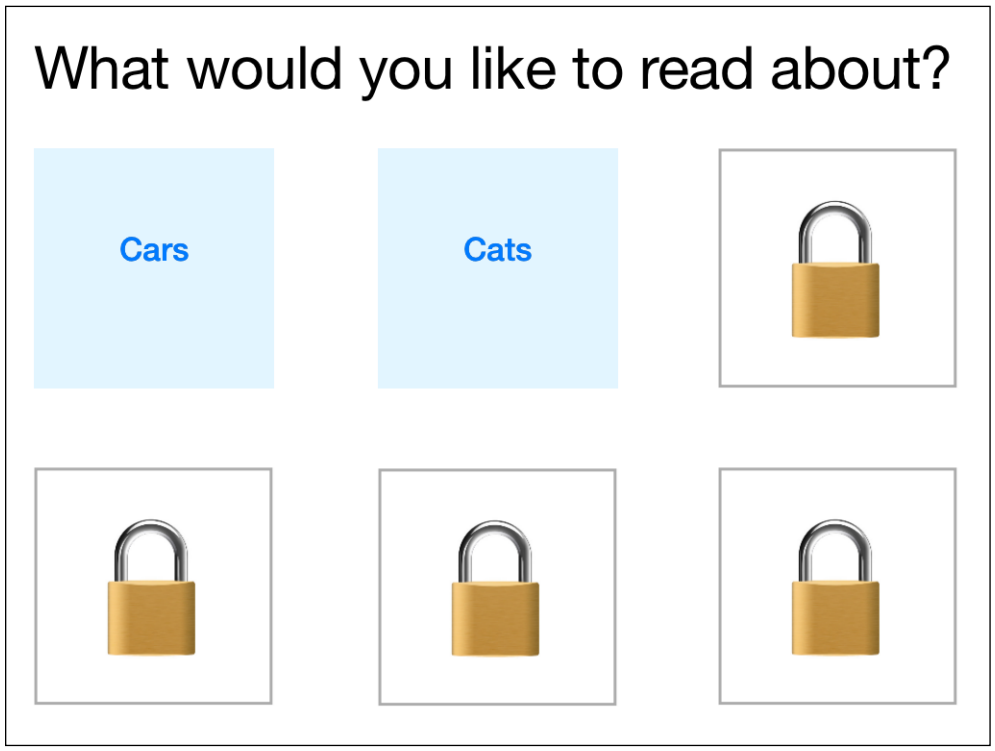

Figure 1: Topic selection screen

After the user selects a topic, they are presented with the text screen shown in Figure 2. The text screen allows the user to read the passage of text, before selecting “Done" to move onto the next screen or “Quit" to exit the application. If the text length exceeds the length of the screen, the user is able to swipe up or down to scroll further into the text. There is no time limit for the user to read the text.

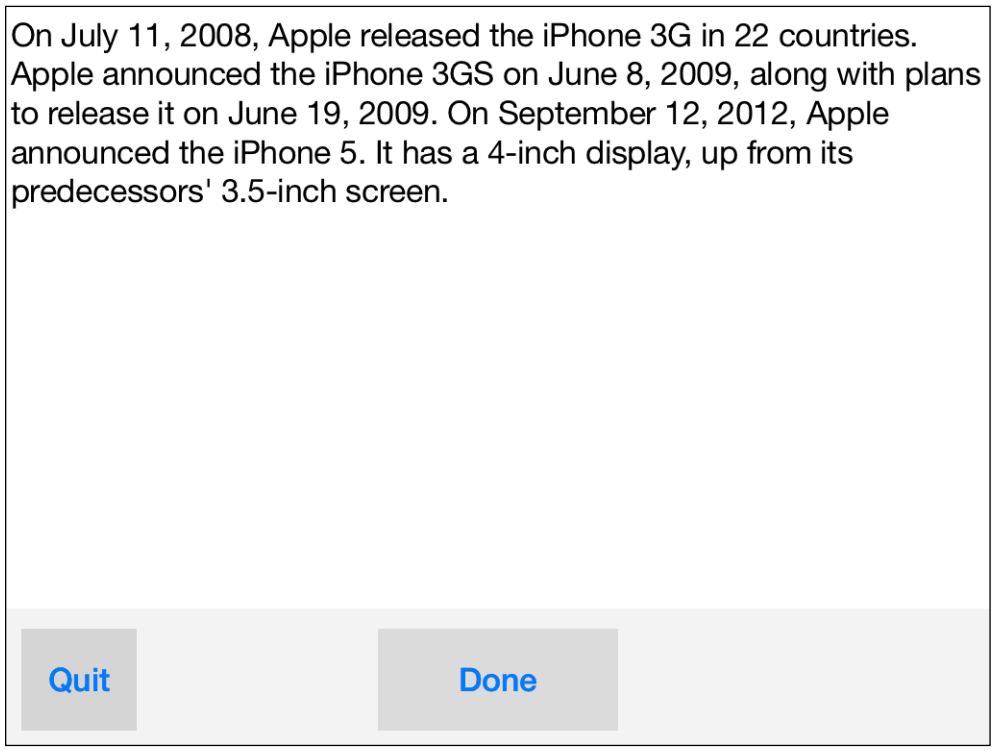

Figure 2: Text screen

After the user clicks “Done" on the text screen they are presented with the question screen found in Figures 3 and 4. The question screen presents a series of questions related to the previous passage. Each question has four potential answers. Only one answer is correct, and the remaining three answers are incorrect. When a user answering a question correctly by taping on the correct response, that response will be highlighted in green as shown in Figure 3. When a user answers a question incorrectly, the incorrect response will be highlighted in red, and the correct response will be highlighted in green, as shown in Figure 4. The top of the question screen allows the user to see how many questions they have remaining, as well as keep track of which questions they have answered correctly by the list of check marks (with correctly answered questions switching from a grey to green check mark).

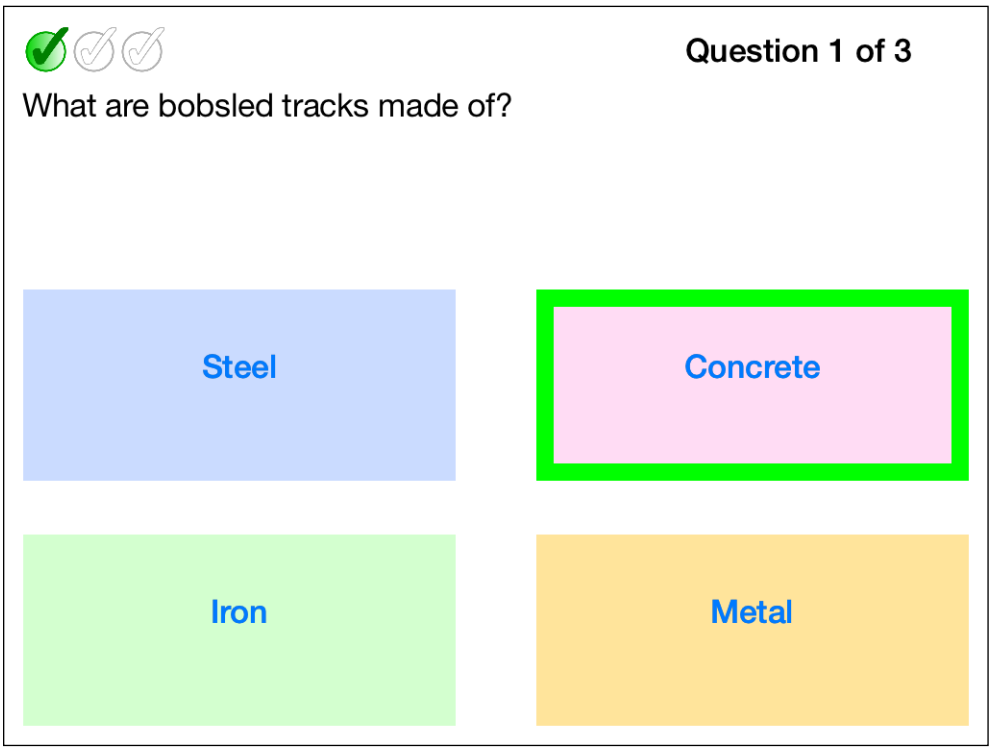

Figure 3: Correct answer

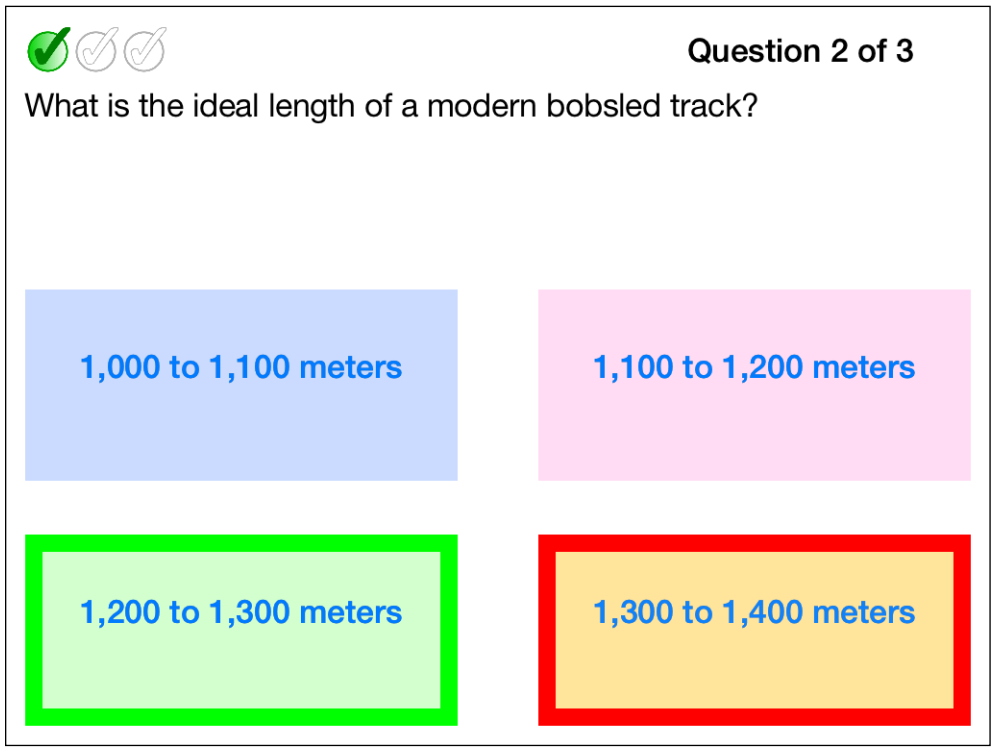

Figure 4: Incorrect answer

Once the user has completed the set of questions associated with the passage, they are forwarded to the reward screen shown in Figure 5. The reward screen keeps track of how many perfect scores the user has accumulated. A perfect score is achieved when a user answers every question associated with a passage of text correctly. The reward screen also keeps track of how many perfect scores are needed to "unlock" another passage topic option in the topic selection screen. Three, four, five and six options are made available to the user on the topic selection screen after achieving two, five, ten and eighteen perfect scores respectively.

When a user has achieved a perfect score in the question screen, before the reward screen is displayed a large gold check mark and "perfect score" text is briefly flashed on the screen. Similarly, when a perfect score results in a topic unlock, before the reward screen is displayed a large unlock icon and "option unlocked" is briefly flashed on the screen.

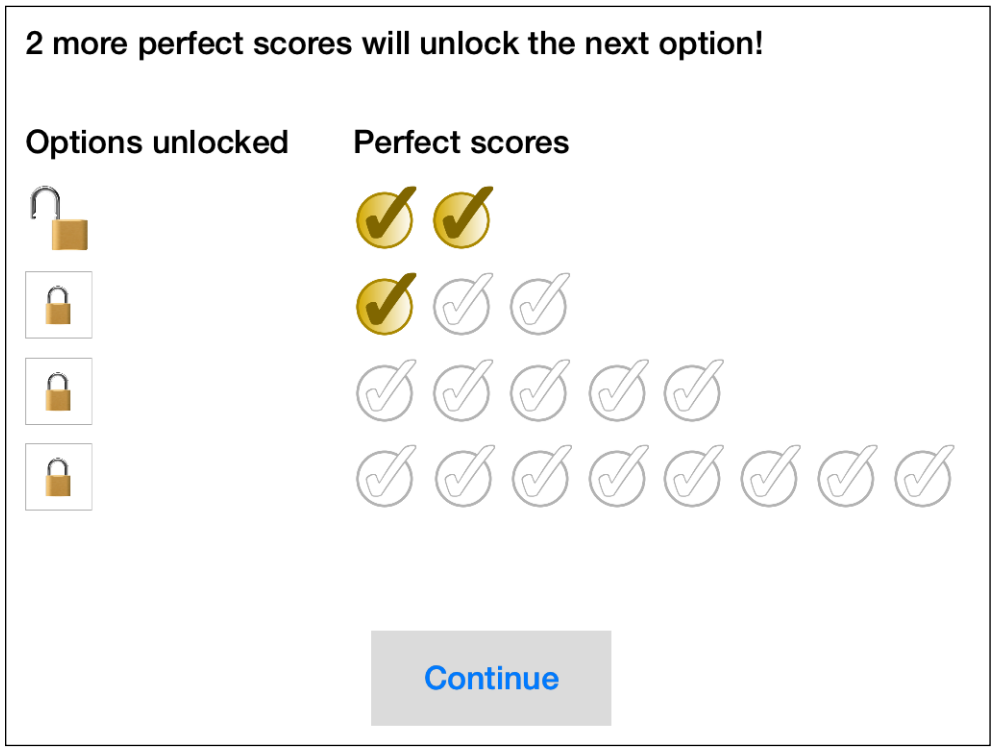

Figure 5: Reward screen

If the user did not receive a perfect score during the question screen, the question generation strategy is presented to the user via a set of consecutive screens shown in Figures 6, 7 and 8.

The motivation screen shown in Figure 6 is meant to show the user that the question generation strategy is effective. This screen randomly cycles through ten research results showing the question generation strategy to be effective.

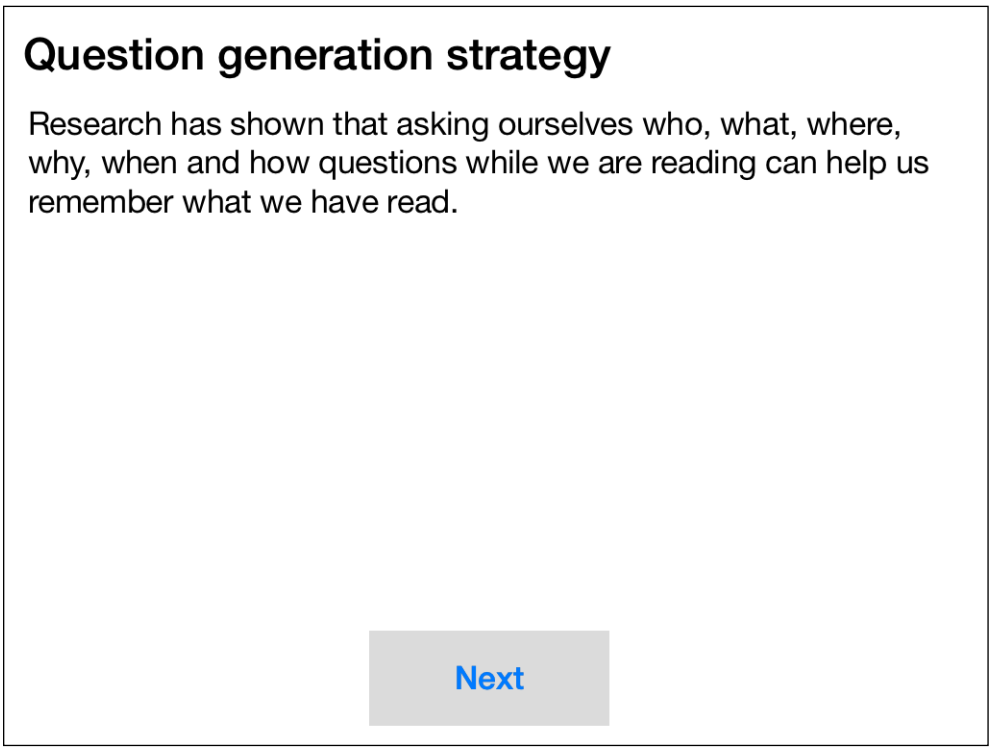

Figure 6: Motivation screen

After clicking “Next" on the motivation screen, the user is presented with an example screen as in Figure 7, demonstrating the question generation strategy, by presenting a passage of text and associated questions that could be derived from the passage of text. The questions are all either who, what, when, where, why or how questions, based on the type of question the user *most recently* answered incorrectly. The example is randomly selected from a set of five possible examples associated with each type of question.

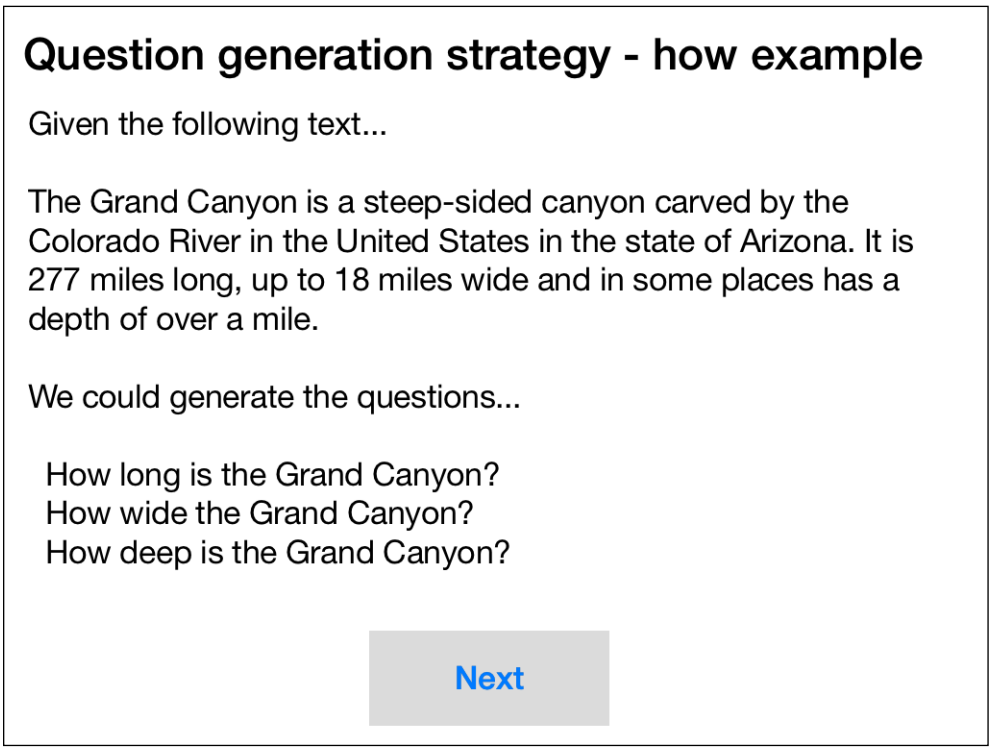

Figure 7: Example screen

After the user clicks “Next" in the example screen, the user is presented with the tactics screen shown in Figure 8. The tactics screen presents a tactic to help the user apply the question generation strategy, i.e. a method of carrying out the question generation strategy. The tactics screen randomly presents one of four possible tactics for carrying out the question generation strategy.

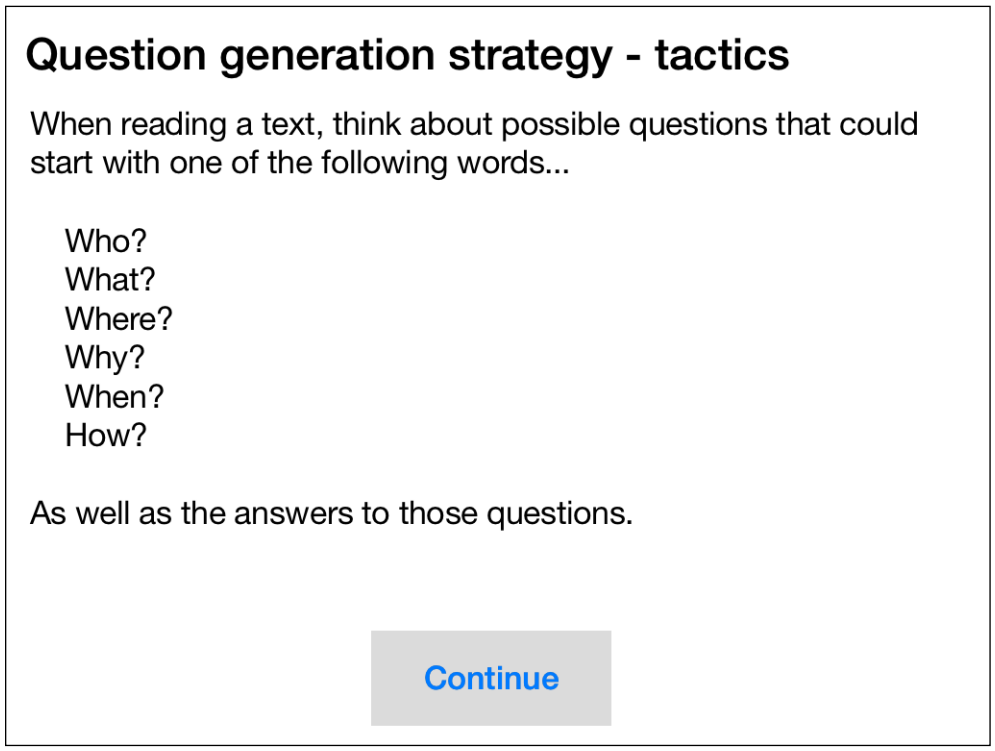

Figure 8: Tactic screen

After the user clicks “Next" in the tactics screen, the user is presented with the topic selection screen again. The application cycles continuously in this way until the user quits the application.

While the user is not made explicitly aware of this process, the application adjusts the difficulty of the passages and associated questions as the user proceeds, based on the user’s performance. The user is able to advance through 7 levels of passages and associated questions. The user begins by receiving level 1 passages and questions. If the user achieves two perfect scores in a row at their current level, then the user advances a level. If the user answers less than 50% of questions correctly three times, then the user returns to the lower level. However, the user will stay in level 1 or level 7, even if they meet the criteria for reversion or advancement, respectively.

The user’s current level determines which passages of text are made available for them to select at the topic selection screen. Each level is made up of 40 different possible texts and topics. The topics

presented to the user at the topic selection screen are random. However, the user will not be presented with the same text options at a given level again until it is no longer possible to present text options that the user has not already read.

Each level has an increasing number of questions associated with each passage in that level. The passages in each level are given an increasing word count range, and a decreasing Flesch-Kincaid score range. The Flesch-Kincaid score ranks the difficulty of a passage of text using metrics such as the total number of sentences, total number of syllables, and total number of words (Kincaid, Aagard, O'Hara & Cottrell, 1981). The levels and associated question count, word count range, and Flesch-Kincaid score range are shown in Table 1.

Level	Questions	Word count range	Flesch-Kincaid score range
1	1	0-20	90-100
2	2	20-40	80-90
3	3	40-60	70-80
4	4	60-80	60-70
5	5	80-100	50-60
6	6	100-140	30-50
7	7	140-180	0-30

Table 1: Passage levels

Creating the 280 passages and questions was a substantial undertaking performed by the first author and freelancers. The passages themselves were, with a few exceptions, taken from Wikipedia articles. If the text from the Wikipedia article did not conform to the required Flesch-Kincaid score, but was relatively close to the required score, words and sentences were altered to ensure that it fit to the desired

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

score. In the case of levels 1-2, it was very difficult to find text on Wikipedia with the required Flesch-Kincaid score. As a result, many passages at these levels were created from scratch.

None of the questions require the user to make an inference based on the text to answer them correctly. Every correct answer is directly presented in the text itself (e.g. a date, a person's name).

We made an effort to select diverse topics (including cultural and gender diversity), although no measure of this diversity was taken. Passage topics included areas such as pop culture (e.g. television, movies, celebrities, musicians), history (e.g. war, the history of nations), science (e.g. biology, chemistry) and others. As an example of cultural and gender diversity, music and musician-related topics spanned several genres with male and female musicians represented.

An effort was also made to ensure an equal balance of who, what, when, where, why and how question types. In the case of levels 1-3, the question types occurred across all passages in equal number. In the case of levels 4-7, it became unwieldy to ensure an equal portion of question types, but no question type was represented over 25% more than the other question types.

Design Approach

Four key elements were woven together in the design:

- Question generation strategy
- Dynamic difficulty adjustment
- Gamification
- Experiential learning

The question generation strategy taught by the application is meant to give the user a metacognitive strategy to improve their reading comprehension (Cohen, 1983; Rosenshine, 1997). Metacognitive

strategies include three components: declarative (“knowing what"), procedural (“knowing how"), and conditional (“knowing why") (Paris, Lipson & Wixson, 1983). Teaching a metacognitive strategy can be done by providing explicit answers for these what, how and why questions (Carrell, 1998). In analogy with this decomposition, we *declare* the name of the strategy (“question generation") in the application. The screens which motivate the strategy provide the *conditional* component, and the concrete examples and tactics for applying the strategy provide the *procedural* component.

Dynamic difficulty adjustment is incorporated into the level system. That two perfect scores in a row are required to move up a level is meant to make it relatively difficult to move into the upper levels by random chance, and that three less than 50% scores in a row are required to move down a level is meant to make moving down a level rare.

Though dynamic difficulty adjustment was intended primarily to increase engagement, the related concept of *flow* may also be encouraged. Nakamura and Csíkszentmihályi (2009) describe flow as a subjective experience that seamlessly unfolds from moment to moment. Csíkszentmihályi (1997) models flow as a balance between perceived opportunities and skills, with the current model of flow shown in Figure 9 having apathy experienced when the perceived challenges and skills are below the user’s average levels, and flow experienced when the challenges and skills are above the user’s average levels (i.e. the stretching of existing skills). By balancing the challenge level to the user’s performance, we may also expect that the user experiences a sensation of flow while using the application. However, we also note by this model that if the balance isn’t achieved we may expect anxiety, apathy, worry or boredom on the part of the users.

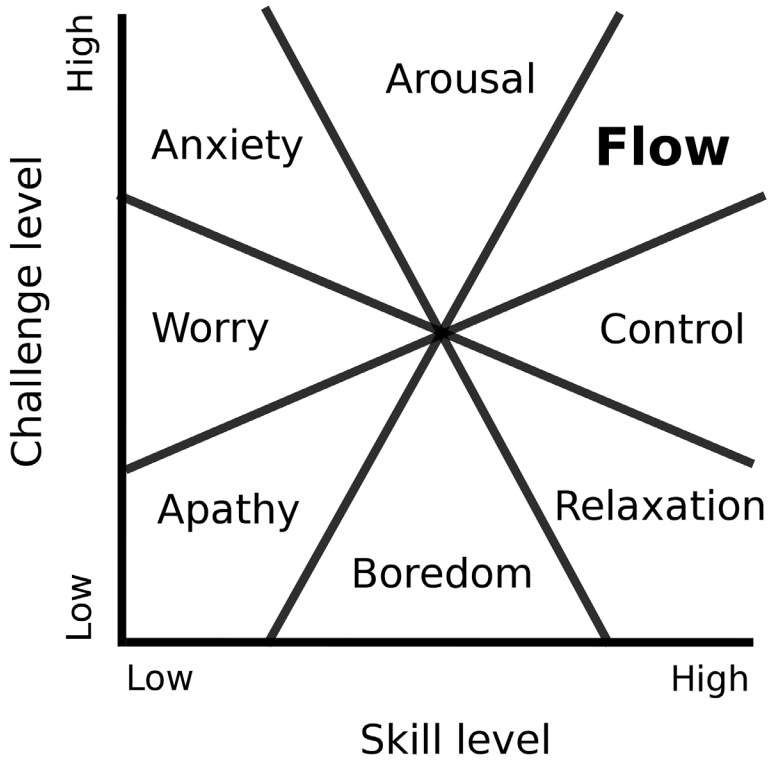

Figure 9: Csíkszentmihályi model of flow

Gamification to increase engagement is facilitated by the following design elements:

- **Badges** - green and gold check marks are given as rewards for successfully answering questions.
- **Levels** - the user has the ability to implicitly proceed through different levels of difficulty depending on their performance.
- **Short, medium, and long term goals** - earning an individual green check mark is a short-term goal, earning a gold check mark is a medium-term goal, and earning enough gold check marks to unlock the next option is a long-term goal.

Distinguishing between short-term goals with green check marks, medium term goals with gold check marks, and long-term goals with an unlocked lock graphic creates a visual layering of goals that

take increasing amounts of time to complete but that come with increasing rewards, a key feature of many games (Dorling & McCaffery, 2012).

Experiential learning is the process of learning through reflection on experience, and can be characterized by a cycle of active experimentation, concrete experience, reflective observation, and abstract conceptualization (Kolb, 1985). The cyclical nature of the application from experimentation and concrete experience (passage and question screens), to reflective observation (results, rewards) to abstract conceptualization (question generation strategy screens) is modeled as such in an attempt to facilitate experiential learning.

Control Application

The control application works by cycling only between the text screen in Figure 2, and the question screen in Figure 3 and Figure 4. No opportunity is given to select a text topic; instead a text is randomly selected from one of the level 7 texts. Level 7 texts were chosen because level 7 texts should be appropriate for university-level readers. The application ensures they will not receive the same text again until all other texts have been used. No reward presentation screens or reward collection screens are displayed. The only gamification that is included is a green check mark upon receiving an individual correct answer. No screens related to the question generation strategy are displayed. The application therefore simply cycles between a randomly selected level 7 text and related questions to allow the user to practice reading comprehension.

Metrics

Both applications recorded metrics such as the level of each passage the user read (all level 7 in the case of the control application), and the total number of passages read.

Experiment Design

McMaster University students were invited to participate in a study measuring improvements in reading comprehension using iPad applications. Participants were recruited using department-wide e-mails to multiple departments in diverse areas of study, and posts to a diverse collection of subject-, activity-, and club-specific Facebook groups.

Participants received a $10 Tim Hortons gift card as compensation for their time and motivation for their participation in the study. The study was approved by the McMaster University Research Ethics Board.

Experiment session protocol

The following protocol was followed with each experiment participant. The protocol refers to the pre-experiment questionnaire in Section 3.4, the usability survey in Section 3.5, and the post-experiment questionnaire in Section 3.6.

The sessions took place in a classroom and a meeting room at McMaster University and were conducted by Browne. The classroom contained desks and chairs facing each other in a circle. The meeting room contained a large table surrounded by chairs.

The sessions took place over a period of 2 weeks. Participants were able to select a convenient time slot. Sessions took place with anywhere from 1 to 7 participants at a time. Sessions took approximately 1 hour to complete.

Half of the participants were given the experiment application, and half of the participants were given the control application. All participants in any single session were given the same application, allowing participants to discuss their experience with that application with their peers. All participants used an iPad Mini tablet device during the sessions. All participants used their version of the application for approximately 30 minutes, as participants were told at the 30-minute mark that they could “now move

on to the post-study practice sheet, but could finish completing the current passage and questions if desired", rather than abruptly cutting them off from the application. Note that 30 minutes of instruction in other reading comprehension strategies has produced significant improvements (Gambrell & Bales, 1986).

The following procedure was used during each session:

1. The participants were told the goal of the experiment.
2. The rest of the experiment procedure was outlined for the participants.
3. The participants completed a paper copy of the pre-experiment questionnaire.
4. The participants completed a reading comprehension practice sheet.
5. The participants used the iPad application for approximately 30 minutes.
6. The participants completed a reading comprehension practice sheet.
7. The participants completed the usability survey.
8. The participants completed the post-experiment questionnaire.
9. The participants were encouraged to discuss their thoughts on the application.

A total of 24 participants used the experiment application (the experiment group), and a total of 24 participants used the control application (the control group).

Two different reading comprehension practice sheets were developed, with a best effort to make them approximately the same level of difficulty. In order to control for a differing level of difficulty between the practice sheets, the practice sheets were alternated within each group of participants (experiment and control). Within each group, half the participants did one practice sheet at the start of

the session and the other at the end of the session, and the other half completed them in the reverse order.

It's reasonable to suspect that if someone simply practiced reading a passage and answering related questions for a period of time, that their score on a reading comprehension practice sheet could go up simply due to practice. The control application was developed for the control group to ensure that any improvement noticed in the experiment group wasn't simply caused by additional practice with reading comprehension passages and questions.

Quantitative observations

Quantitative observations were recorded using the reading comprehension practice sheets to test the participants throughout the experiment session. Each reading comprehension practice sheet contained three passages and associated questions: a level 3 passage and 3 associated questions, a level 5 passage and 5 associated questions, and a level 7 passage and 7 associated questions. Each question had 4 possible answers, and only one answer was correct in each case. Each practice sheet had a maximum score of 15. The passages and questions used for the practice sheets were not included in the iPad applications.

Qualitative observations

After each session, a casual verbal discussion with the participants was used to elicit further insights into the effectiveness of the applications. Observations from these discussions were recorded in writing. Verbal expressions, reactions, and comments made by the participants during the sessions were also recorded in writing as study data. Qualitative observations of user perception of each application were also recorded with the usability survey in Section 3.5.

Pre-experiment questionnaire

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

The following information was gathered with the pre-experiment questionnaire:

- Gender (Male/Female)
- Age
- Handedness (Right/Left)
- Years of study completed at the University level
- Year of study in current program
- Current program of study The participants were also asked to rate their reading ability from 1 to 5: 1 is “not well at all" and 5 is “I can read perfectly well", and asked to rate their ability to use the iPad from 1 to 5, where 1 is “not well at all" and 5 is “I can use the iPad perfectly well".

Usability survey

The participants were asked to rate how much they agree (Likert scale) with the following statements:

- **S1** The app was easy to use.
- **S2** It was easy to learn how to use this app.
- **S3** I enjoyed using this app.
- **S4** The iPad was comfortable to hold while using the app.
- **S5** The app helped me to improve my reading comprehension.
- **S6** I found the app to be useful.
- **S7** I would tell other people to use this app.
- **S8** The touchscreen finger gestures required to use the app felt natural.

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

- **S9** I liked the app's graphics.
- **S10** I liked the app's voices / sound.
- **S11** The app kept me totally absorbed.
- **S12** The app held my attention.
- **S13** The app excited my curiosity.
- **S14** The app aroused my imagination.
- **S15** The app was fun.
- **S16** The app was intrinsically interesting.
- **S17** The app was engaging.
- **S18** Using the app was interesting in itself.
- **S19** Using the app was fun.
- **S20** I thought of other things while using the app.
- **S21** I felt curious while using the app.
- **S22** I was in control of the app that I was using.
- **S23** I was entirely absorbed in using the app.

The participants could choose from: strongly disagree, disagree, somewhat disagree, neutral, somewhat agree, agree, and strongly agree. For analysis purposes, these descriptions were assigned numeric values 1-7 from strongly disagree to strongly agree.

For brevity's sake we will refer to this survey as the usability survey, but we note that a group of questions are intended to measure aspects of usability, the next group is intended to measure

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

engagement, and a final group is intended to measure flow. Statements S1-S10 are intended to measure usability and are identical to those used in our prior study of the effectiveness of tablet software to teach adult literacy skills. Statements S11-S17 are modeled closely after those used in a survey to measure engagement in a prior study by Webster and Ho (1997). Statements S18-S23 are modeled closely after those used in a survey to measure flow in a prior study Choi and Kim (2004). As a result of combining multiple different survey instruments, some survey questions are very similar (e.g. S15 and S19).

Post-experiment questionnaire

The following questions were asked on the post-experiment questionnaire.

1. Would you prefer to be taught reading comprehension using the iPad app or by some other method? (check one)
2. In the future should people be taught reading comprehension only using the iPad app, only using some other methods, or both? (check one)
3. What did you like about the iPad app? (write below)
4. What didn't you like about the iPad app? (write below)

Results and Discussion

A total of 48 participants took part in the experiment, 24 participants in the experiment group used the experiment application and 24 participants in the control group used the control application. The programs of study reported by the participants in the pre-study questionnaire were wide ranging in both the experiment and control group, to such a degree that each group only contained a few instances of participants from the same area of study. The remaining participant data collected during the pre-

experiment questionnaire is presented in Table 2. We note that reading ability and iPad ability as reported by the participants were closely matched between the groups.

	Experiment group	Control group
Gender		
Women	15	13
Men	9	11
Handedness		
Left	2	2
Right	22	22
Age		
Average	22.9	22
SD	5.8	2.9
Reading ability		
Average	4.5	4.5
SD	0.7	0.8
iPad ability		
Average	4.1	4.2
SD	0.9	0.9
Years of University		
Average	4.3	3.9

SD	3.1	2
Years in current program		
Average	2.8	3
SD	1.1	1.2

Table 2: Participant data

With a sample group of exclusively McMaster University students, the results of this study cannot be extended to the general population. However, given the reasonably random participant selection process, we believe our results are statistically significant for the sampled population of those who came in contact with the recruitment materials. When we talk about results being statistically significant for the population, it is this population we refer to and not the general population.

The practice-sheet results are shown in Figure 10, where no improvement in score was found in the control group but the experiment group did improve their average score. The average performance of the control group went from 11.0 (s=2.359) to 10.875 (s=3.353), and the average performance of the experiment group went from 10.958 (s=2.368) to 12.708 (s=2.579).

We conduct the following analysis of variance hypothesis test at significance level α=0.05. We used the ANOVA caluclator (two-factor ANOVA with repeated measure on one factor) available at www.vassarstats.net.

Null and alternative hypotheses:

1. $H_0;\mu_{exp}=\mu_{con}$

   $H_A;\mu_{exp}\neq\mu_{con}$

2. H_0;$\mu_{pre}=\mu_{post}$

   H_A;$\mu_{pre}\neq\mu_{post}$

3. H_0; an interaction is present

   H_A; an interaction is absent

Test statistic:

We compute F_{BS}, F_{WS}, and $F_{BS\times WS}$ in a 2×2 mixed-design analysis of variance model where the between-subjects variable is the iPad application (either experiment or control) and the within-subjects variable is the practice sheet (either pre-application usage or post-application usage).

Decision rules:

1. If F_{BS} is greater than 4.05, we reject the null hypothesis.

2. If F_{WS} is greater than 4.05, we reject the null hypothesis.

3. If $F_{BS\times WS}$ is greater than 4.05, we reject the null hypothesis.

Note: the chosen significance level implies these critical values.

Computing the test statistic:

We present the results of the ANOVA in summary Table 3, where SS is the sum of squares, df is the degrees of freedom, MS is the mean square, and F is the test statistic.

Source	SS	df	MS	F
Between Subjects	453.24	47		

Factor	19.26	1	19.26	2.04
Error	433.98	46	9.43	
Within Subjects	261.5	48		
Factor	15.84	1	15.84	3.25
Factor	21.09	1	21.09	4.32
Error	224.57	46	4.88	
Total	714.74	95		

Table 3: ANOVA Summary Table

Applying the decision rules:

1. F_{BS}=2.04<4.05, therefore we fail to reject the null hypothesis.

2. F_{WS}=3.25<4.05, therefore we fail to reject the null hypothesis.

3. $F_{BS \times WS}$=4.32>4.05, therefore we **reject the null hypothesis**.

The statistically significant interaction suggests that the experiment application successfully taught the experiment group participants the question generation strategy, and that participants improved their reading comprehension skill as a result (at least temporarily).

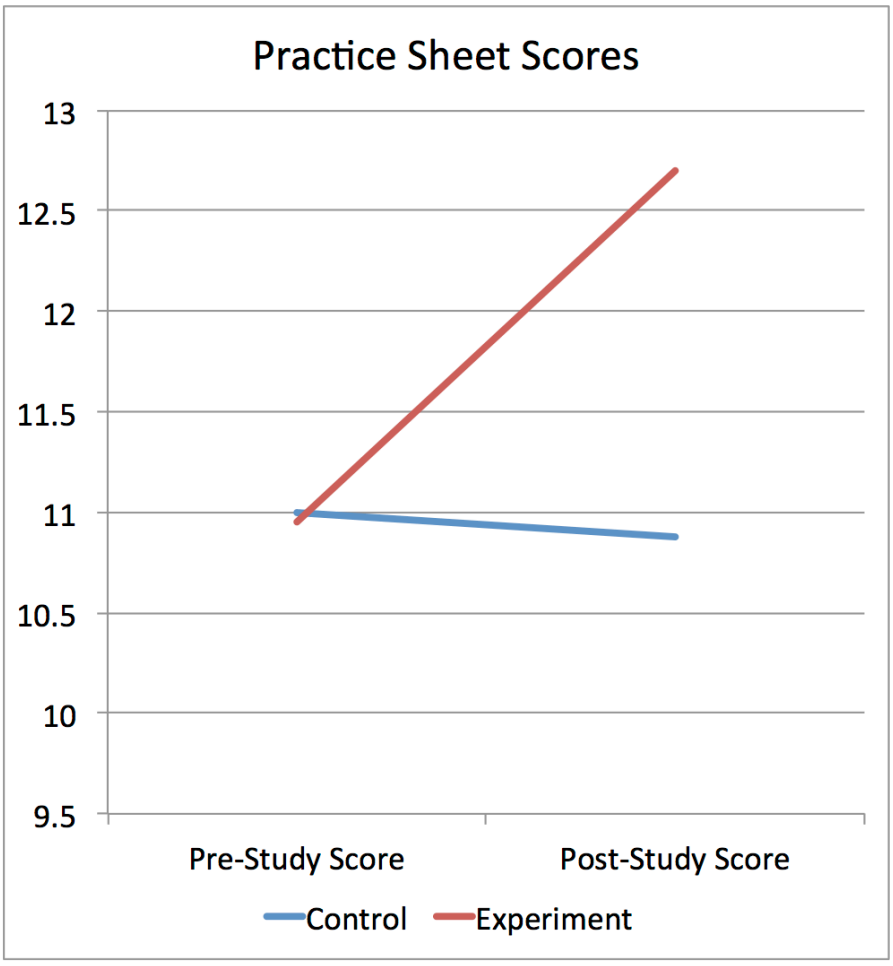

Figure 10: Practice sheet scores

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

The result of the usability survey are presented in Figure 11. The results do not show any substantial difference between the applications, though we note that in all but two statements (S9 and S22) the experiment application received higher results. We noted in Section 3.5 that the usability survey was comprised of three sections, S1-S10 focusing on usability, S11-S17 focusing on engagement, and S18-S23 focusing on flow. Within each group of questions, we sum the average result of each question (e.g. summing the average of S1-S10), divide it by the total highest average possible (e.g. divide by 70 in the case of S1-S10), and multiply the result by 100 to give a rough score. We present these results in Figure 12, where again we note that the experiment application has received higher results, but not significantly higher results as was thought possible.

Figure 11: Usability survey results

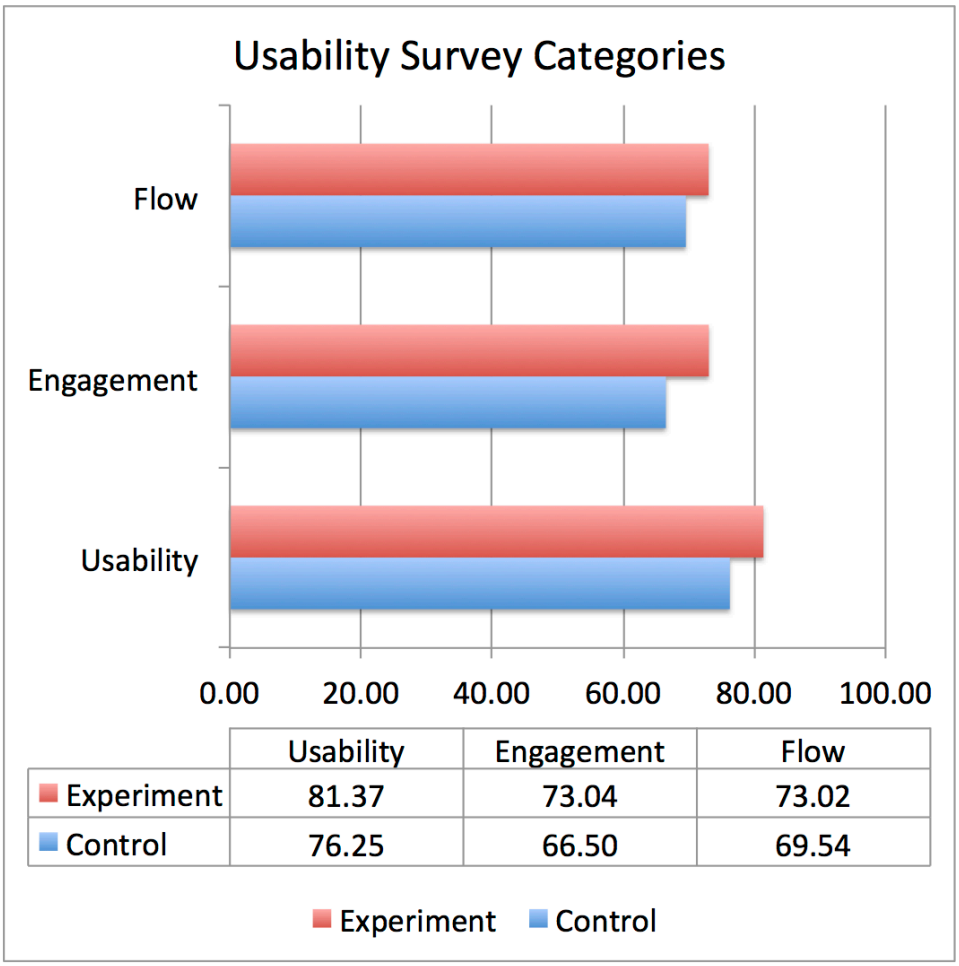

Figure 12: Usability survey categories

In Table 4 we present the results of the post-experiment question asking participants to choose their preference between the iPad application or some other method for learning reading comprehension.

	iPad	**Some other method**
Control	15	9
Experiment	18	6

Table 4: Preferred learning method

In Table 5 we present the results of the post-experiment question asking participants to recommend how reading comprehension should be taught in the future (iPad application, some other

other method, or both). The participants in both groups showed a very strong preference for using both the iPad application and some other method going forward. Again, these results do not provide evidence that the experiment application is preferable to the control application as was thought possible.

	iPad app only	**Some other method**	**Both**
Control	0	1	23
Experiment	0	0	24

Table 5: Preferred future learning method

The free form written post-experiment questionnaire questions and informal discussions with participants offer a reasonable explanation for why the experiment application did not result in a significantly more engaging and preferred experience.

In the case of positive feedback for both applications, participants noted that they were “easy to use".

In the case of positive feedback for the experiment application, participants noted the layered rewards and would even use the word engaging to describe what they were feeling, “It was engaging since you had to improve in order to get to the next level." The ability to select a topic to read was also singled out for praise in the experiment application. Some participants also noted that they were aware of the question generation strategy that the application had taught them, and expressed a feeling that it could help them.

The negative feedback for the control application included having to read passages for which the participant was not interested, that the application was “boring", and a belief that it wasn’t effective at improving reading comprehension. All of these reactions are things that we anticipated.

The negative feedback for the experiment application was roughly divided into two groups. One group of participants would either give little to no negative feedback, where the negative feedback could

be described as “wanting more". These participants were generally happy with the application, but they desired things such as more topics, images to go along with the passages, and more complexity.

The other group of participants found the application to be frustrating. While advancing through levels 1-4 was generally pretty easy for most participants, advancing to levels 5,6 and 7 was more difficult. If a single question in the set of questions associated with a passage was answered incorrectly, a perfect score would not be achieved and the participant could not advance. This requirement for a perfect score to advance, and a general feeling that they needed to advance, was noted by several participants as a cause of frustration. For example, one participant gave the feedback, “If you got one question wrong, you could not unlock the next round." Another cause of frustration was that once participants reached level 7, they were not able to advance further. Participants described this as a lack of finality. For example, one participant gave the feedback, “Use of perfect scores to unlock progression diminished sense of attainability of a final goal."

As a result of this feedback, we believe that the application’s implementation of dynamic difficulty adjustment was mismatched with the expectations of these users, resulting in frustration in some cases. Though participants were not explicitly or purposely made aware of the different levels of difficulty, participants quickly figured out that these different levels of difficulty existed based on the number of questions they were presented. Once participants were aware of these levels, they naturally used them as a yardstick to measure their progress, and became frustrated as a result at the lack of advancement.

In this sense the design of the application was not in keeping with the purpose of dynamic difficulty adjustment. If a game adjusts its difficulty dynamically, strictly speaking it has implemented dynamic difficulty adjustment. But if the goals are not also adjusted to reflect the new level of difficulty, or if the goals are associated with adjustment towards higher levels of difficulty, the user may see the increase or decrease in difficulty as an advancement or setback in their progress.

A further literature review revealed work by Andrew Rollings and Ernest Adams (2003) which emphasized the importance of hiding the existence of dynamic difficulty adjustment from the player for the technique to work as intended. The failure of our application to hide dynamic difficulty adjustment may have been the reason the dynamic difficulty adjustment implementation was poorly received by some participants.

It should be noted that some of the frustration that participants expressed appeared to be motivational. A few participants would talk about how they were frustrated while using the application, but then pump their fists after achieving a perfect score and advancing to another level. However for some participants this frustration appeared to cross over from an enjoyable level into upset, disappointment, and a deeper frustration that was no longer motivating.

In Table 6 we present data on the total number of passages read for each group. All but one participant in the experiment group was able to reach level 7. The minimum number of passages it took to reach level 7 was 6, the maximum was 16, with an average of 8.9 (s=2.9).

	Min	**Max**	**Avg**	**Stdev**
Control	7	34	16.4	6.7
Experiment	9	24	17.2	3.7

Table 6: Total passages read

Conclusion

Our main question was: “Can reading comprehension performance and user engagement be improved by teaching the question generation strategy via tablet software incorporating dynamic difficulty adjustment and gamification strategies.” We can answer with high confidence that reading

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

comprehension can be improved, at least for the population sampled over the short interval of the experiment.

While the results cannot be generalized to other populations, they do provide valuable insights into reading comprehension educational software design. Due to the design, the question generation strategy was inherently linked to difficulty adjustment and gamification, but we interpret the lack of significant difference in preference, usability, engagement, and flow between the applications, and the fact that all but one participant in the experiment group reached level 7, to mean that it was the instruction of the question generation strategy that resulted in improved practice sheet scores within the experiment group, rather than the dynamic difficulty adjustment or gamification features of the experiment application. In fact, the feedback suggests that our application design frustrated some participants.

The participants' feedback could be incorporated by shortening the length of time participants spent with the application, so participants would not reach level 7, or, alternatively, providing a reward for reaching level 7, to signify that the user had "completed" the application (akin to "beating" a game). Some users may also prefer a level promotion system not based on perfect scores, but another metric such as average scores at the current level.

Participants' feedback also suggests that the implemented gamification techniques (e.g., layered rewards (Hamari & Eranti, 2011) and gradually increasing difficultly levels (Keller, Bless, Blomann & Kleinböhl, 2011) are effective in the domain of reading comprehension tablet software. However, in our experience, dynamic difficulty adjustment is a difficult system to implement well, requiring the maintenance of a balance between the level of challenge and each user's ability, with rewards to maintain engagement. We recommend an iterative design and test approach.

Our work and results can be beneficial to literacy educational technology creators and literacy researchers. The results of this work should give designers of educational literacy software increased

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

confidence that they can teach reading comprehension strategies in their applications effectively. An abundance of literature exists covering various reading comprehension strategies and how to teach them to an individual or a class of a students, but very few studies exist in which the strategy is taught via software. Given the encouraging results of this work, and the economic benefits of delivering education via software, we believe there is an opportunity for researchers to develop new applications which teach different metacognitive strategies, perhaps using different design features (e.g. videos). The effectiveness and viability of metacognitive strategies taught using digital platforms can likely be explored extensively given the abundance of research on the original strategies themselves.

In future work, we hope to:

- Improve the dynamic difficulty adjustment implementation of the experiment application to increase user engagement.
- Test the experiment application (or an improved version of it) with adult-literacy-centre clients to investigate whether it can improve their reading comprehension skills.
- Compare the experiment application to non-tablet software with adult-literacy-centre clients, measuring improvements in reading comprehension.
- Test whether other reading comprehension strategies can be delivered via tablet software, beyond the question generation strategy, and measuring their relative effectiveness.
- Design an application using dynamic difficulty adjustment to assess reading skill level, and comparing it to current methods used to triage students and tailor reading programs.
- Measure the effectiveness of an expanded application over an extended period of time.

Volume 19, Number 2: Winter 2018
ISSN: 1535-0975

The highest priority is to incorporate the insights gained from this study and to test this new application with adult-literacy-centre clients, since their needs originally motivated this work.

References

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. *Handbook of reading research*, *1*(353), V394.

Barr-Telford, L., Nault, F., & Pignal, J. (2005). Building on our competencies: Canadian results of the International Adult Literacy and Skills Survey. *Statistics Canada. Available at: www. statcan. ca/bsolc/english/bsolc*.

Barth, A. E., & Elleman, A. (2017). Evaluating the Impact of a Multistrategy Inference Intervention for Middle-Grade Struggling Readers. *Language, Speech & Hearing Services In Schools*, 48(1), 31-41.

Bean, T. W., & Steenwyk, F. L. (1984). The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. *Journal of Literacy Research*, *16*(4), 297-306.

Bell, N., & Lindamood, P. (1991). *Visualizing and verbalizing: For language comprehension and thinking*. Paso Robles, CA: Academy of Reading Publications.

Bereiter, C., & Bird, M. (1985). Use of thinking aloud in identification and teaching of reading comprehension strategies. *Cognition and instruction*, *2*(2), 131-156.

Bishop, D. V., & Adams, C. (1990). A prospective study of the relationship between specific language impairment, phonological disorders and reading retardation. *Journal of Child Psychology and Psychiatry*, 31(7), 1027-1050.

Bower, J. V., & Rutson-Griffiths, A. (2016). The relationship between the use of spaced repetition software with a TOEIC word list and TOEIC score gains. *Computer Assisted Language Learning,* 29(7), 1238-1248.

Cain, K., & Oakhill, J. (1998). Comprehension skill and inference-making ability: Issues of

causality. *Reading and spelling: Development and disorders*, 329-342.

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of educational psychology*, *96*(4), 671.

Carrell, P. L. (1998). Can reading strategies be successfully taught?. *Australian Review of Applied Linguistics*, *21*, 1-20.

Choi, D., & Kim, J. (2004). Why people continue to play online games: In search of critical design factors to increase customer loyalty to online contents. *CyberPsychology & behavior*, *7*(1), 11-24.

Cohen, R. (1983). Self-generated questions as an aid to reading comprehension. *The Reading Teacher*, 36(8), 770-775.

Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life*. Basic Books.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, *19*(4), 450-466.

Deterding, S. (2011, May). Situated motivational affordances of game elements: A conceptual model. In *Gamification: Using game design elements in non-gaming contexts, a workshop at chi*.

Dorling, A., & McCaffery, F. (2012, May). The gamification of SPICE. In *International Conference on Software Process Improvement and Capability Determination* (pp. 295-301). Springer Berlin Heidelberg.

Gambrell, L. B., & Bales, R. J. (1986). Mental imagery and the comprehension-monitoring performance of fourth-and fifth-grade poor readers. *Reading Research Quarterly*, 454-464.

Hamari, J., & Eranti, V. (2011, September). Framework for designing and evaluating game

achievements. In *Proc. DiGRA* (Vol. 115, pp. 122-134).

Hunicke, R. (2005, June). The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology* (pp. 429-433). ACM.

Jetton, T. L., & Dole, J. A. (Eds.). (2004). *Adolescent literacy research and practice*. Guilford Publications.

Johnson, P. (1981). Effects on reading comprehension of language complexity and cultural background of a text. *TESOL quarterly*, *15*(2), 169-181.

Keller, J., Bless, H., Blomann, F., & Kleinböhl, D. (2011). Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology*, *47*(4), 849-852.

Kincaid, J. P., Aagard, J. A., O'Hara, J. W., & Cottrell, L. K. (1981). Computer readability editing system. *IEEE Transactions on Professional Communication*, (1), 38-42.

Kolb, D. A. (1985). Learning styles inventory. *McBer, Boston, MA*, b8.

Kutner, M., Greenberg, E., & Baer, J. (2006). A First Look at the Literacy of America's Adults in the 21st Century. NCES 2006-470. *National Center for Education Statistics*.

Lapp, D., Flood, J., & Ranck-Buhr, W. (1995). Using multiple text formats to explore scientific phenomena in middle school classrooms. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *11*(2), 173-186.

Long, E. (2001). *Patterns of Participation in Canadian Literacy and Upgrading Programs: Results of a National Follow-Up Study*. ABC Canada Literacy Foundation, 333 King Street East, 3rd Floor, Toronto, Ontario M5A 4N2, Canada.

McGuinness Publishing. (2011, April, 1). *Apple's Steve Wozniak's Vision Of The Future Of*

Computing - Showing The Knowledge Navigator concepts [Video file]. Retrived from https://www.youtube.com/watch?v=eRNRg589S2c

McNamara, D. S. (Ed.). (2012). *Reading comprehension strategies: Theories, interventions, and technologies*. Psychology Press.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 222-233.

Missura, O. (2015). *Dynamic Difficulty Adjustment* (Doctoral dissertation, Bonn, Rheinische Friedrich-Wilhelms-Universität Bonn, Diss., 2015).

Nakamura, J., & Csikszentmihalyi, M. (2009). Flow theory and research. *Handbook of positive psychology*, 195-206.

Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic reader. *Contemporary educational psychology*, *8*(3), 293-316.

Paris, S. G., Wasik, B., & Turner, J. C. (1991). The development of strategic readers.

Pearson, P. D., Hansen, J., & Gordon, C. (1979). The effect of background knowledge on young children's comprehension of explicit and implicit information. *Journal of Literacy Research*, *11*(3), 201-209.

Peng, P., & Fuchs, D. (2017). A Randomized Control Trial of Working Memory Training With and Without Strategy Instruction. *Journal Of Learning Disabilities*, 50(1), 62-80.

Rollings, A., & Adams, E. (2003). *Andrew Rollings and Ernest Adams on game design*. New Riders.

Rosenshine, B. (1997). The case for explicit, teacher-led, cognitive strategy instruction. *MF Graves (Chair), What sort of comprehension strategy instruction should schools provide*.

Rushton, S. P., Eitelgeorge, J., & Zickafoose, R. (2003). Connecting Brian Cambourne's conditions of learning theory to brain/mind principles: Implications for early childhood educators. *Early Childhood Education Journal*, *31*(1), 11-21.

Shang, H. F., & Chang-Chien, I. J. (2010). The Effect of Self-Questioning Strategy on EFL Learners' Reading Comprehension Development. *International Journal of Learning*, *17*(2).

Smith, T. (2016). EXPONENTIAL READING GROWTH Celebrating Digital ELA Curriculum Success. *Tech & Learning*, 37(5), 22-33.

Sousa, I., & Oakhill, J. (1996). Do levels of interest have an effect on children's comprehension monitoring performance?. *British Journal of Educational Psychology*, *66*(4), 471-482.

Stanovich, K. E., & Siegel, L. S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology*, *86*(1), 24.

Toukonen, K. (2011). *The Dynamic Electronic Textbook: Enhancing the Student's Learning Experience* (Doctoral dissertation, Kent State University).

Vernon, J. A., Trujillo, A., Rosenbaum, S., & DeBuono, B. (2007). Low health literacy: Implications for national health policy. Washington, DC: Department of Health Policy, School of Public Health and Health Services, The George Washington University.

Webster, J., & Ho, H. (1997). Audience engagement in multimedia presentations. *ACM SIGMIS Database*, *28*(2), 63-77.

Wong, B. Y., & Jones, W. (1982). Increasing metacomprehension in learning disabled and normally achieving students through self-questioning training. *Learning Disability Quarterly*, *5*(3), 228-240.

Yin, H. S., Johnson, M., Mendelsohn, A. L., Abrams, M. A., Sanders, L. M., & Dreyer, B. P.

(2009). The health literacy of parents in the United States: a nationally representative study. *Pediatrics*, *124*(Supplement 3), S289-S298.

Yuill, N., Oakhill, J., & Parkin, A. (1989). Working memory, comprehension ability and the resolution of text anomaly. *British journal of psychology*, *80*(3), 351-361.