Gamification of Mobile Educational Software

Gamification of Mobile Educational Software

By
Kevin Browne B.Sc. (Hons), M.Sc.
A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree
Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY(2016) McMaster University
COMPUTING AND SOFTWARE Hamilton, Ontario

TITLE: Gamification of Mobile Educational Software

AUTHOR: Kevin Browne B.Sc. (Hons)(McMaster University), M.Sc. (McMaster University)

SUPERVISOR: Dr. Christopher Anand

NUMBER OF PAGES: xx, 207

LEGAL DISCLAIMER: This is an academic research report. I, my supervisor, defence committee, and university, make no claim as to the fitness for any purpose, and accept no direct or indirect liability for the use of algorithms, findings, or recommendations in this thesis.

Lay Abstract

The overall theme of this thesis is the study of how gamification can be used to make mobile educational software engaging and effective as a learning tool. The research presented in this thesis focuses on the design and testing of software intended to teach introductory computer science and literacy concepts. The hypothesis guiding this work is that mobile educational software can be made engaging and educationally effective by incorporating game design elements. Through four studies we affirm our hypothesis, and document the relative success of various gamification techniques in different contexts. We make design suggestions for software creators, such as providing corrective feedback to the user. We discuss common themes that emerged across the studies, including how to best use educational software. Finally, as avenues for future work, we suggest investigating the impromptu social effects of using tablet software in a classroom, and the development of a usability testing platform.

Abstract

The overall theme of this thesis is the study of incorporating gamification design approaches in the creation of mobile educational software. The research presented in this thesis focuses on the design and testing of software created to teach introductory computer science and literacy concepts to post-secondary and adult learners. A study testing the relative effectiveness and subjective user enjoyment of different interfaces for a mobile game is also included in this thesis, as the results of the study led to the primary research objectives investigated in further studies.

Our primary research objective was to investigate whether using gamification design approaches to mobile educational software could result in student engagement and learning. Our central hypothesis is that gamification design approaches can be used to create engaging and educationally effective mobile educational software. Our secondary research objective is to determine *how* mobile educational software can be made more or less engaging and educationally effective through gamification design approaches, by trying different approaches, testing the resulting applications, and reporting the findings.

Three studies were conducted based on these objectives, one study to teach various computer science concepts to students in a first year computer science course with iPad applications, and two studies which used iPad applications to teach punctuation and homonyms, and improve reading comprehension. The studies document the design of the applications, and provide analysis and conclusions based on the results of testing.

Through the results of these studies we affirm our hypothesis. We make design suggestions for software creators, such as providing corrective feedback to the user. We discuss common themes that emerged across the studies, including how to best use educational software. Finally, as avenues for future work, we suggest investigating the impromptu social effects of using tablet software in a classroom, and the development of a usability testing platform.

Acknowledgments

I would like to thank my supervisor Dr. Christopher Anand for his support, guidance and feedback. In particular Dr. Anand encouraged me to become involved in activities such as developing computing and software outreach materials, conducting events on campus where industry experts and recent graduates would speak with students, and working as a sessional lecturer. Though these activities might not be required to complete a PhD, having the freedom to do them during my PhD significantly influenced and enriched my research, career and life directions.

I would like to thank my supervisory committee (Dr. Melina Head, Dr. Spencer Smith, Dr. Wolfram Kahl) for their support, guidance and feedback.

I would like to thank NSERC and OGS for their financial support.

I would like to thank my parents, partner, family, and friends for their support and encouragement.

Contents

List of Figures

List of Tables

List of all Abbreviations and Symbols

HCI - Human-Computer Interaction

Declaration of Academic Achievement

This document is presented as a "sandwich thesis" in accordance with the Guide for the Preparation of Theses at McMaster University. The work presented in Chapters 2, 3, 4 and 5 comprise four separate papers that are unified by the overall research objectives of this thesis. The research objective of investigating gamification as a design approach to mobile educational software was motivated by the study presented in Chapter 2, and investigated by the studies presented in Chapters 3, 4 and 5. My contributions to the four studies as first author are described in this section. As of August 2015, the papers in Chapters 2, 3, and 4 have been published in peer-reviewed journals or conference proceedings. The paper presented in Chapter 5 will be submitted to a peer-reviewed journal or conference.

Chapter 2 *K. Browne, and C. Anand, An empirical evaluation of user interfaces for a mobile video game, Entertainment Computing 3.1 (2012) 1-10. (doi:10.1016/j.entcom.2011.06.001)*

Contributions included:

- Primary responsibility for identifying the research problem.
- Primary responsibility for obtaining study clearance from the McMaster University Research Ethics Board.
- Primary responsibility for the design and implementation of the iOS application developed for the study.
- Primary responsibility for the experiment design.
- Primary responsibility for conducting the usability study experiment with participants and associated data collection.
- Primary responsibility for data analysis and interpretation.
- Primary responsibility for writing and submission of the paper to journal.
- Primary responsibility for alterations and communication in response to feedback from peer reviewers.

Chapter 3 *K. Browne and C. Anand, Gamification and serious game approaches for introductory computer science tablet software, Proceedings of the First International Conference on Gameful Design, Research, and Applications. ACM, 2013.*

Contributions included:

- Primary responsibility for obtaining study clearance from the McMaster University Research Ethics Board.
- Primary responsibility for conducting the usability study experiment with participants and associated data collection.
- Primary responsibility for data analysis and interpretation.
- Primary responsibility for writing and submission of the paper to journal.
- Primary responsibility for alterations and communication in response to feedback from peer reviewers.
- Shared responsibility with Christopher Anand for identifying the research problem.
- Shared responsibility with Christopher Anand for the experiment design.
- Shared responsibility with Christopher Anand and summer undergraduate students for design of the iPad applications.

In this paper, the identification of the research problem and experiment design was shared equally with Christopher Anand. The implementation of the iPad applications in this study was done by summer students working for Christopher Anand. The design of the applications was a shared responsibility between Christopher Anand, Kevin Browne, and the undergraduate students. Teaching assistants working in the classes in which the experiment in this study took place assisted in conducting the usability study experiment.

Chapter 4 *K. Browne, C. Anand, E. Gosse, Gamification and serious game approaches for adult literacy tablet software, Entertainment Computing 5.3 (2014) 135146. (doi:10.1016/j.entcom.2014.04.003)*

Contributions included:

- Primary responsibility for identifying the research problem.
- Primary responsibility for making contact with and organizing study in coordination with the Brant Skills Centre.
- Primary responsibility for obtaining study clearance from the McMaster University Research Ethics Board.
- Primary responsibility for the design and implementation of the iPad applications developed for the study.
- Primary responsibility for the experiment design.
- Primary responsibility for conducting the usability study experiment with participants and associated data collection.
- Primary responsibility for data analysis and interpretation.
- Primary responsibility for writing and submission of the paper to journal.
- Primary responsibility for alterations and communication in response to feedback from peer reviewers.

Chapter 5 *K. Browne, C. Anand, An empirical evaluation of reading comprehension tablet software utilizing the question generation strategy (unpublished as of August 2015)*

- Primary responsibility for identifying the research problem.
- Primary responsibility for obtaining study clearance from the McMaster University Research Ethics Board.
- Primary responsibility for the design and implementation of the iPad applications developed for the study.
- Primary responsibility for the experiment design.
- Primary responsibility for conducting the usability study experiment with participants and associated data collection.
- Primary responsibility for data analysis and interpretation.
- Primary responsibility for writing of the paper.

Chapter 1

Introduction

This thesis is structured as a sandwich thesis. In the introduction we present the overall themes, motivations, background information, research objectives, hypothesis and provide a summary of the included studies. In the subsequent chapters we present the individual studies conducted as part of this work, verbatim to their presentation in journals and conferences for those studies which have been previously published. We include amendments in each chapter for clarifications, corrections and further analysis in response to feedback received from reviewers. In the conclusion we provide analysis of the work done and suggest future research directions, in both cases pertaining to the work as a whole rather than rehashing conclusions and future research directions already covered in the individual studies.

In the introduction we overview the theme of the work done in Section 1.1, discuss our motivations for conducting the work in Section 1.2, and provide relevant background information in Section 1.3. In Section 1.4 we discuss our philosophical approach to this work in relation to the different types of work that can be conducted in this area. In Section 1.5 we state the thesis research objectives and in Section 1.6 we discuss our hypothesis. Finally in Section 1.7 we summarize the individual papers presented in the subsequent chapters.

1.1 Overview

The overall theme of this thesis is the study of incorporating gamification design approaches into the creation of mobile educational software. The studies presented in this thesis focuses on the design, development and testing of software created to teach introductory computer science and literacy concepts to post-secondary and adult learners. An article documenting a study testing

the relative effectiveness and subjective user enjoyment of different interfaces for a mobile game is also included in this thesis, as the results of the study led to the primary research objectives investigated in further studies.

1.2 Motivations

There were several factors which motivated us to study the gamification design approaches in mobile educational software. Motivations arose both organically from the results of our own work, and from broader movements occurring in games research, education and technology. In Section 1.2.1 we review how motivations arose organically from the results of our own work, and in Section 1.2.2 we review motivations that arose from broader movements occurring in games research, education and technology.

1.2.1 Prior Work

Prior to conducting the studies in Chapters 3, 4 and 5, we conducted a study to investigate the effectiveness and enjoyability of a mobile game playable with different interfaces, and developed an educational game to teach computer science to elementary school students as part of an outreach activity. The results of this prior work raised questions about the potential for mobile educational software incorporating gamification elements, and the success of this prior work gave us confidence that we could meaningfully investigate these questions.

The main motivation was the outcome of the study discussed in Chapter 2. In this study we designed, implemented and tested an iPod Touch game developed. The game was a "scroll shooter" style game with enemy spaceships scrolling down the screen and firing projectiles at a player controlled ship with the ability to fire projectiles to destroy enemy ships. The game was built with three interfaces: accelerometer (tilt and shake), touchscreen gesture (swipe and tap) and touchscreen buttons. We tested the three interfaces with 36 participants, and were able to obtain interesting data relating to interface preference and performance. This study gave us confidence that we could meaningfully investigate the effectiveness and enjoyment of different game design approaches relative to one another.

The motivation to study educational technology came from our experience with the McMaster Software Outreach program. We developed an application aimed at getting grade 7 and 8 students interested in studying computer science through a classroom presentation. The author created a game of pong for Mac OS X that could be played with two Wiimotes. The game starts off

"broken" with the rules of pong missing (i.e. the ball flies through the paddle) and the students were encouraged to make the game work as expected by implementing the rules in Objective-C. As the game is rebuilt rule-by-rule, students would come up to the front of the class and play each version of the game with the Wiimotes. While this project was not intended as a research activity, we observed during presentations in classrooms that students showed a high level of focus, understanding and participation. We believed it was primarily the game that caused these positive (though informal) observations, as other outreach activities had not elicited such a reaction before.

The high level of engagement and understanding we witnessed with the outreach activity was something we wished to investigate with the same formalism and rigor as the scroll shooter study—both to confirm our suspicions that the inclusion of a game or game elements in learning activities could lead to such a reaction, and to determine what specific game design approaches would work best. It was at this point that we began to search the literature, and investigate broader trends in education and technology, to ensure that our own suspicions, interests and motivations would be of value to the broader research community.

1.2.2 Tablet Computers, Gamification and Education

We investigated the expected growth in sales of table computer devices and their use in educational environments, to ensure tablet computers would be widely used in the problem domain of education, and to find potential problems and questions about the usage of tablet computers in education. We reviewed literature to determine whether these questions had been investigated, and for what problem domains within education, to ensure our work would be novel and valuable. We reviewed literature relating to gamification, again to ensure that our studies would be valuable and novel within this area as well. We also investigated the importance of the education problem domains of introductory computer science and adult literacy to ensure demand and utility for studies done in this area. After this literature review we concluded that studing the gamification of tablet computer educational software for introductory computer science and adult literacy would indeed be of value to the research community and society.

When this work began in 2010, shipments of tablet computers like the iPad were expected to grow from 19.5 million units in 2010 to 208 million units in 2014, according to Gartner Inc. media analysts [GAR10]. Though tablet sales slowed in 2014, they surpassed this expectation and are projected to grow to 468 million by 2017[Gar15].

At the same time school boards such as the Hamilton-Wentworth District School Board and the Los Angeles Unified School District are making major investments into tablet devices; in the case of Los Angeles Unified School District $50 million dollars was spent to equip 30,000 students with an iPad[Pec13, Leo13]. There is apprehension amongst parents, educators and students as to the effectiveness and necessity of these devices:

"If the object was to allow kids to have a free iPad to go to Facebook and to engage in social media, that has been accomplished. But what it has to do with better learning, I don't know." - Diane Ravitch, author of Reign of Error: The Hoax of the Privatization Movement and the Danger to Americas Public Schools[Leo13]

"I just see a mad rush from a lot of districts right now just to say, 'We bought iPads. Now what?' That worries me." - Michael Horn, executive director of the Clayton Christensen Institute's education program[Leo13]

This apprehension wasn't too surprising given the novelty of the technology. In 2010-2012 there was a lack of strong peer-reviewed research into the effectiveness of tablet computers as learning devices. There was some promising early research, for example, Houghton Mifflin Harcourt conducted a year-long pilot study in a middle school in Riverside, California in which an iPad application was used to increase student performance by 20% in algebra compared to peers who used textbooks[Che12]. Other studies also suggested tablets could be useful and effective in the classroom[HY12], with other studies focusing on the effectiveness of a particular software design[CG11]. The success of these early studies and need for more studies to determine how best to design and use educational tablet software provided justification for our desires research direction.

Though using game design elements in software and education isn't new[Kap12], around 2010-2011 the notion of *gamification* reached a new level of maturity, rigor and interest with work by Deterding et al to define the concept[DDKN11]. Though gamification was widely recognized for being a trending topic garnering much hype in the period after 2010, whether it really *worked* or not, and what that meant, was a question still open for investigation[HKS14]. Indeed, some of the same worries about the hype surrounding tablet software in the classroom were echoed in concerns about the effectiveness of Gamification. One of the world's leading IT research and analysis companies, Gartner has warned that 80% of all gamification apps will fail to meet their objectives due to poor design[Gar12].

Though the problem domain of introductory computer science applications was a natural offshoot of our discipline and activities in the computing & software outreach program, the growing "learn to code" movement was also

a motivating factor[She14, Kal15]. The shift from low-skill, low-wage jobs to high-skill, high-wage jobs due to automation of routine tasks by computer software[A+10], and its effects on employment in the City of Hamilton which has seen manufacturing jobs drop by over 30,000 in a decade[Ruf13]. Though promoting and enhancing computer science education may be beneficial to society, as we conducted our work we discovered that beyond *digital literacy*, traditional *english literacy* was a barrier to many entering the knowledge economy workforce.

The International Adult Literacy and Skill Survey found that 48% of the Canadian population over 16, representing 12 million people, have literacy skills below the level required for coping with the increasing demands of a knowledge economy[BTNP05, DMT05]. People with low literacy skills are more likely to have lower rates of employment and to work in low-skill jobs[BTNP05]. In the USA, a 2005 report by the National Assessment of Adult Literacy found that 93 million people lack the literacy skills required to complete the education or job training required by current and future jobs[KGB05]. In addition to keeping workers out of the knowledge economy, the economic and health impacts of low literacy skills have been extensively documented by international and national organizations [Hig08, Lit01]. In the USA, an estimated $106 to $238 billion dollars in annual health-care costs can be attributed to poor health knowledge and behaviour resulting from low literacy[VTRB07, YJM+09]. For these reasons we chose to investigate educational tablet software to teach literacy skills to adult learners.

Demand for research into education delivery exists, as usage of scientific research as a basis for decisions regarding education delivery was ratified into law in the USA with the No Child Left Behind Act of 2001[Cho03]. Meta-analysis of the literature regarding instruction techniques has found that using computer software which involve the manipulation of symbolic artifacts as a learning tool showed a percentile gain of 43 points in student achievement [Mar98]. This give us reason to believe that educational software can be educationally effective, and will be in demand.

Tablet computers and gamification represented, and still represent, exciting new possibilities to transform and improve education. However, it was clear that in both cases a gap in the peer-reviewed literature existed around these topics, in terms of how best to design and use tablet educational software and software incorporating gamification. The problem domains of computer science and adult literacy were particularly appealing given our own discipline, the growing shift towards knowledge economy jobs, and the potential benefits to society.

1.3 Background and Key Concepts

The work presented in this thesis intersects with gamification, education and human-computer interaction. At the outset of this research we discovered that each of these fields was a subject of significant academic analysis.

In this section, we review the most relevant literature, each with what we deem the appropriate depth, focusing on definitions and ideas useful for understanding and contextualizing our work at the time it began. In Section 1.3.1 we provide definitions for mobile devices and tablet computers. In Section 1.3.2 we review ideas relating to gamification, in Section 1.3.3 we review ideas relating to education, and in Section 1.3.4 we review ideas relating to human-computer interaction. Most of these topics contain great breadth, and in some cases such as gamification, continue to evolve rapidly.

1.3.1 Mobile Devices and Tablet Computers

The terms "mobile device" and "tablet computer" do not have definitions that are as clear, precise or commonly accepted as other terms in computing such as "firewall" or "encryption"[Fri10]. It's even debated whether tablet computers should be considered mobile devices, for example Facebook CEO Mark Zuckerberg proclaimed in 2010 that the iPad was not a mobile device[Chi10]. In the absence of clear, precise and commonly accepted definitions, we provide definitions for the purposes of our work that we believe to be reasonable.

We define a *mobile device* to be a computing device small enough to be used or operated while being held in one hand or both hands.

We define a *tablet computer* to be a battery-powered mobile device with a touch screen display between 7 and 13 inches, capable of recognizing finger gestures (including gestures which require multiple fingers).

We used Apple products (iPod Touch, iPad, and iPad Mini) in our studies. We consider the iPod Touch to be a mobile device, and the iPad and iPad Mini to be tablet computers.

1.3.2 Gamification

We discuss the definition and usage of gamification, because each of the studies in Chapters 3, 4 and 5 explore and investigate the usage of gamification as a design approach.

Gamification is a design approach which has produced promising increases in engagement in several different contexts. Gamification can be defined as the "usage of game design elements to motivate user behaviour in

non-game contexts"[Det11]. An example would be the location-based social network Foursquare which rewards users for checking in to their current location with points and allows the user to become the "mayor" of that location. Other examples include startup company ZamZee[Zam12] that was able to use a gamification-based app to increase physical activity in children by 60%, and the Greater Washington Give Day who were able to use gamification to drive two million dollars in fundraising in one day[Bar12].

Gamification is recognized as a relatively new concept, and as such, alternative definitions are still being proposed[HH12, Det11]. For example Huotari and Hamari[HH12] define gamification as "a process of enhancing a service with affordances for gameful experiences in order to support user's overall value creation." However, some are critical of gamification as a term, suggesting instead that it only amounts to putting a scoring system on top of a non-game activity[Nic12].

The common element across definitions of gamification is the usage of game design elements in (subjectively defined) non-game contexts. The usage may not necessarily be associated with motivation. The usage of game mechanics in non-game contexts may simply be used to make an activity subjectively more enjoyable, rather than strictly speaking more *motivational* in some meaningful sense such as increasing a user's physical activity.

Game design elements may include a range of mechanics and features such as rewards, levels, time limits, collection keeping, bidding, points, role-playing, tile-placing, and multiplayer. Gamification may involve a simple scoring system, however the complexity of the achievement systems can become quite sophisticated[HE11], and the game design elements used as part of a gamification design approach may go beyond simple scoring systems.

Serious Games

Related to gamification is the concept of a serious game, a design approach used in the applications designed for the studies in Chapters 3 and 4.

Serious games can be defined as "digital games used for purposes other than mere entertainment"[SJB07]. Serious games are used in a variety of domains, from rehabilitation[RMR10] to education[MC05].

Serious Games vs. Gamification

The studies in Chapters 3 and 4 feature applications that one could reasonably argue are examples of serious games. In the case of the study in Chapter 3, we make a distinction that some applications are serious games and some are

not, and in Chapter 4 we explicitly design one application as a serious game and one application is designed to utilize gamification design concepts (but we do not consider it a game).

The distinction between serious games and gamification is considered unclear in the literature[McC12], since both involve the usage of game design elements in a non-game context. The difference is that serious games incorporate enough game design elements to warrant calling the software a *game*. In some sense this makes serious games a subset of software designed with game design elements (i.e. gamification). The boundary between gamification and serious games is recognized as being blurry[Det11], based on subjective and social factors.

Games

Though the boundary between gamification and serious games may be blurry, we provide a definition for what we believe constitutes a *game* to provide further clarity about our own views.

We follow the definition of a game as prescribed by Ernest Adams in Fundamentals of Game Design: "Games are a type of play activity, conducted in the context of a pretended reality, in which participant(s) try to achieve at least one arbitrary, nontrivial goal by acting in accordance with the rules."[Ada13]

Though we follow this definition of a game, we recognize that numerous other overlapping definitions exist[SZ04, Cos08, Cla87, Mar01].

Tablet Educational Software with Gamification Design

We searched for any existing studies examining tablet educational software utilizing gamification, given that all the applications built for the studies in Chapters 3, 4 and 5 are themselves instances of tablet educational software utilizing gamification.

At the time we began this work, there had been a few studies looking at educational-tablet game software. Wattanatchariya et al. developed a game called "Drop Donuts" for teaching wind and gravity concepts and reported on survey results from 12 university student participants[WCD11]. Feng Yan developed an iPad application "A Sunny Day" to educate children with autism, which features mini-game and puzzle-game elements, and tested it with autistic children and their parents[Yan11]. Yan's work is especially valuable because of attention to application-design elements, to testing of the interface with target users, and reporting on what did and didn't work and why. For example, their

analysis suggests that the software should have had clearer objectives and provided rewards more quickly after completing tasks. The paper concludes that the app offers a cost-effective therapeutic approach relative to existing methods.

1.3.3 Education

When deciding to study education-related problem domains, we reviewed the literature relating to education technology, educational tablet software, and literacy education. This literature review influenced the design of the applications used in our studies

There is a wealth of literature that attempts to understand educational practices. The topics include areas such as education policy, pedagogy, and curriculum development, and the work draws from intersecting areas such as sociology and psychology. We chose to focus on a few ideas relevant to our research focus

In this section we review education theory topics that influenced our studies and application designs.

Cambourne's Conditions for Learning

To design the adult literacy applications in Chapter 4, we searched for literature relevant to teaching literacy and found Brian Cambourne's eight Conditions of Learning.

Brian Cambourne developed eight Conditions of Learning for literacy development[REZ03], and the conditions are summarized as follows:

1. **Immersion** - learners need to be immersed and constantly saturated in that which is to be learned.

2. **Demonstration** - learners need to receive many stimulating demonstrations of desired outcomes.

3. **Engagement** - learners must be engaged in the learning process while being immersed in the learning environment and viewing demonstrations.

4. **Expectations** - learners are influenced by expectations, which are powerful shapers of behavior.

5. **Responsibility** - learners need to make their own decisions about when, how, and what "bits" to learn.

6. **Employment** - learners need time and opportunity to use and practice new learning in realistic ways.

7. **Approximation** - learners must be free to approximate desired study, as mistakes are essential for learning to occur.

8. **Response** - learners must receive relevant, appropriate, timely, non-threatening feedback.

Educational Tablet Software

We were interested in how these conditions could be facilitated via *tablet software* specifically, and we found work by Kayne Toukonen helpful in guiding application design decisions.

Kayne Toukonen looked at how the features of a tablet device could facilitate learning in the form of what he termed *dynamic electronic textbooks*[Tou11]. Toukonen pays particular attention to Brian Cambourne's eight Conditions of Learning[REZ03] that Cambourne felt necessary for the acquisition of language, connecting each of these eight conditions to the possible software features enabled by tablets. For example, it is suggested that the learning condition *engagement* can be facilitated by using virtual worlds in tablet applications. Cambourne's conditions of learning and Toukonen's thoughts on how they can be facilitated with tablet software guided some of the design decisions in the applications built for our studies.

Metacognitive Reading Strategies

The final study discussed in Chapter 5 attempted to improve reading comprehension via the teaching (through tablet software) of a specific technique called the question generation strategy. The question generation strategy is a type of metacognitive reading strategy intended to improve reading comprehension.

Metacognitive reading strategies involve having the reader consciously reflect about what they have read, with many experimental results showing that readers can improve their performance by using these metacognitive strategies[SCC10, WJ82, BB85]. By teaching these metacognitive strategies to struggling readers, even in a single learning session, the performance of the readers has been shown to improve[GB86, BB85]. Examples of metacognitive reading strategies include attempting to visualize the text[BL91], or attempting to summarize the text[BS84]. Our study utilized the *question generation* reading comprehension strategy, which involves having the learner generate and answer questions in the process of reading the text[Coh83, Ros97].

Experiential learning

Experiential learning also influenced some of the application designs, in particular the reading comprehension application discussed in Chapter 5.

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[Kol85]. Though experiential learning has several definitions, properties and models[Gen90], all of our applications involve the user applying concepts in an interactive setting.

1.3.4 Human-Computer Interaction

All of the studies presented in this thesis involve the designing and testing of software used by humans. This places the work in the field of Human-Computer Interaction (HCI), a multi-disciplinary field with a distinct terminology useful for discussing the ideas and results in this thesis. We discuss human-computer interaction and some related terminology used in our studies to accurately discuss and help contextualize the work presented.

Human-Computer Interaction (HCI) is defined as "a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them"[HBC+92]. Within HCI is the notion of usability: "the usability of a system is the capability in human functional terms to be used easily and effectively by the specified range of users, given specified training and user support, to fulfill the specified range of tasks, within the specified range of scenarios"[Sha91]. Our work was informed by the ideas and approaches discussed in "Designing the User Interface" by Shneiderman and Plaisant[SB10].

There are many ways to investigate the usability of software. For example, expert reviews depend on having interface and/or problem domain experts formally evaluate software by using it themselves, perhaps checking for specific criteria such as conformance to a list of design heuristics. Another method to investigate the usability of software may be data logging and reporting while the software is in active use.

One key method for evaluating usability is *usability testing*, which involves evaluating software by testing it on users in a formal setting and manner (e.g. using time limits, following a process). The setting could be a sophisticated usability lab with one-way mirrors, or it could be a classroom.

Usability testing may take into account factors such as:

- Time to learn - how long does it take to learn to use the software?
- Speed - how quickly can users perform tasks?
- Errors - at what rate do users commit errors?
- Subjective satisfaction - how much did users enjoy the application?
- Effectiveness - how effective was the software in achieving the desired result(s)?

In the case of educational software, effectiveness may be defined as how well users are able to learn target concepts using the software.

Other factors that may be tested for are engagement and flow. Engagement and flow, perhaps even more so than subjective satisfaction, are states of mind whose definitions are more elusive and less precise than other factors such as error rate. One definition of engagement is, "a positive, fulfilling, work-related state of mind that is characterized by vigor, dedication, and absorption."[SSGRB02] One definition of flow is "an extremely enjoyable experience, where an individual engages in an on-line game activity with total involvement, enjoyment, control, concentration and intrinsic interest"[HL04].

A usability test may involve the software recording aspects of performance such as task-completion and error rates. A usability test may involve survey instruments such as the System Usability Scale (SUS) [BKM08] meant to measure and score usability, or survey instruments designed for more specific factors such as engagement[WH97] and flow[CK04]. A usability test may also involve more informal observations, such as simply observing and documenting participant behavior while using the software, or a Q&A interview with participants after they have used the software.

Applications that are built for testing may be designed around a hierarchy of low-level guidelines and heuristics, mid-level principles and high-level theories and models for usability.

Low-level guidelines give highly specific recommendations for good user interface design practices, as well as warnings against pitfalls. Examples of a guideline or heuristic would be to "use radio buttons for mutually exclusive choices" or to "put a minimal memory load on the user". A criticism of such heuristics is that they are too specific and sometimes inappropriate.

Mid-level principles are used to analyze and compare design alternatives. Principles are less specific than heuristics and tend to be more enduring and widely applicable. An example of principles would be the eight golden rules of user interface design[SB10], which include the rule "strive for consistency". A criticism of mid-level principles is that they may require clarification to be useful.

High-level theories and models are used to describe objects and actions with consistent terminology (for explanation and comprehension), as well as make predictions and prescribe actions. An example of such a high-level model would be Fitt's Law [Mac92], which models human performance in moving to a target area as a function of distance to and size of the target.

Other important high-level user interface theories include direct manipulation and affordance. Direct manipulation interfaces [Shn83] use metaphors for real world objects and actions as the basis for a user interface, which can make them easier for users to learn. Direct manipulation interfaces also allow for quick incremental actions that are easily reversible. Affordance refers to a situation where an object's sensory characteristics intuitively imply its functionality and use [Nor88].

Perhaps in contrast to more mathematical areas of computer science, user interface research involves the application of the scientific method where conjectures are formed, an experiment or test with a hypothesis is executed, and theories are revised and created on the basis of the results. Therefore the applications tested during a usability test may be designed around guidelines, principles and theories, but a usability test may also discover new or provide evidence for existing guidelines, principles and theories.

1.4 Philosophical Approach

As can be appreciated from the previous Section, there are many different types and foci of research in the areas of human-computer interaction, educational software, and gamification, with different objectives leading to different types of outcomes. For example, there is work focused on the definition of ideas and terms[Det11], in contrast to work focused on software design and associated outcomes (positive or negative) obtained after user testing[Yan11].

It is our view that all of this work is valuable in its own right for different reasons. For example, work that focuses on the definition of terms is valuable for giving designers and analysts a common terminology with which to discuss and reason about software and devices. But successful projects need reasonable scoping.

Our work focusses on software design and the analysis of the results of user testing, in order to add empirical support to the literature concerning the value of gamification in our problem domains. As a result of this approach, we do not, for example, dwell on the difference between gamification and serious games or attempt to refine their definitions. As a practical matter, this means that we describe some of our studies as being related to "gamification and serious games" to where any distinctions are not relevant. We do not address the discussion as to exactly what engagement or flow are, but in our final study we choose a reasonable survey instrument to help us measure these properties.

We also do not comment on the value of our chosen type of research relative to other types, as we believe an appreciation for different types of study is important in a research area such as user experience that draws from many different disciplines. However, we do hold that this type of work is a worthwhile addition to the literature and valuable to society who will benefit from better evidence-based educational application designs.

1.5 Objectives

Our primary research objective was to investigate whether applying gamification design approaches to mobile educational software could result in student engagement and learning.

Our secondary research objectives revolve around *how* to design mobile educational software that is engaging and educationally effective, and include:

- Determining what and how specific game design elements are effective, and in what contexts.
- Confirming existing design heuristics and principles for educational and mobile software and gamification.
- Proposing new design heuristics and principles for the gamification of mobile educational software.
- Determining how design approaches based on educational theory can work in concert with gamification design approaches to educational software.
- Investigating how mobile educational software can and should be used as a learning tool, relative to other learning methods and tools.

1.6 Hypothesis

Our central hypothesis is that quantitative experiments can show that gamification design approaches can be used to create engaging and educationally effective mobile educational software.

We are also as interested in secondary research objectives although they are more exploratory and open ended in nature, and expect to find some interesting but mixed results due to the nature of the work and the number of unknowns. We believe that specific gamification design approaches will have varying levels of success in terms of user engagement and educational effectiveness, depending upon the context (type of users, problem domain) and implementation (specifics of the design). We expect that existing design heuristics and principles relating to mobile and educational software and gamification will be re-confirmed, and that some new heuristics and principles will be discovered and supported.

1.7 Summary

We note for the benefit of the reader that there is much overlap between the motivation and background information content presented in this thesis introduction, and the introduction to each individual study.

For the purposes of easier reference within the thesis, we give the individual studies conducted as part of this thesis short names.

Chapters 2 and 3 were published before the draft of this thesis was sent to reviewers for commentary. McMaster Univeristy requires that the original published papers appear as they were originally published in the sandwich thesis. We have therefore placed an additional subsection after the bibliography in these chapters entitled "Amendments". It is in these subsections that we address the valuable feedback from our reviewers regarding these papers.

The first study presented in Chapter 2 is the *Scroll Shooter Study*. As discussed in Section 1.2.1, the results of this study were the key motivator for the research objectives of the subsequent work. In this study we designed and built a "scroll shooter" style video game for the iPad Touch, playable with three different interface styles: touch screen simulated buttons, touch screen gestures and accelerometer. The touch screen simulated button interface was based on a D-pad and button interface found on video game consoles. The touch screen gesture interface was based on swiping and taping gesture movements made on the touch screen. The accelerometer interface was based on tilting and shaking movements applied to the iPod Touch. Our research objective was to

investigate the relative effectiveness and enjoyability of the three interfaces, motivated by the growing mobile game industry. The game was tested with 36 participants, with each participant playing each version of the game. We found a preference for the accelerometer interface. The results with respect to effectiveness (i.e. performance) and enjoyability led us to ask whether we could find similar results with mobile educational software. The most obvious area for us to investigate next was educational software for teaching computer science concepts.

The next study conducted was the *Computer Science App Study* presented in Chapter 3. We designed and built six iPad applications to teach the following introductory computer science concepts: binary search, binary numbers, CPU architecture and assembly language, polynomial graphs, quicksort and Dijkstra's algorithm. Our primary research aims were to investigate whether the applications were effective at teaching these concepts, and whether the applications could increase student satisfaction and engagement. We tested the applications in the introductory computer science course at McMaster University CSM1A3 Computer Based Problem Solving, during the regular course lab sessions over six weeks in the term. A total of 101 students signed up to participate in the study (we allowed students in the course to not participate if desired), with varying numbers of participants in each week of the study (due to the normal variance in class attendance from week to week). Our experiment design compared the educational effectiveness and engagement of the applications against traditional instructional methods (i.e. slide and blackboard based lectures, Q&A). We found that students preferred the applications compared to traditional academic instruction, however we also found that students recommended combined instruction with both traditional methods and the applications in future iterations of the course. Both the applications and traditional instruction led to improved performance, and based on student feedback we suggested that the applications were more suited to practice in applying concepts and traditional methods were more suited to teaching the concepts.

The next study conducted was the *Literacy App Study* presented in Chapter 4. We designed and built three iPad applications to teach literacy skills (punctuation and homophones) to adult clients of the Brant Skills Centre. One of the applications was built with an approach more closely resembling a serious game, and two of the applications were built with an approach more closely resembling gamification of a learning system. A total of 14 participants took part in a test of the homophone application, and a total of 13 participants took part in a test of the punctuation applications. We again compared the applications against traditional instructional methods, and found that most

participants preferred using the applications *in addition* to more traditional instruction. We again found that participants preferred the applications for practicing the concepts learned, and traditional instruction for being taught the concepts themselves. We found that the participants had a range of skill levels that led to frustration with more difficult parts of the applications (or boredom with easier parts of the applications), which suggested investigating a form of dynamic difficulty adjustment as future work.

The final study conducted was the *Reading Comprehension App Study* presented in Chapter 5. We began this study after we had obtained and analyzed the results of the literacy app study and computer science app study. The results of those studies were promising in terms of demonstrating that educational tablet software built with a gamification design approach can be engaging to the point that learners prefer learning with both the software and traditional methods. However, in both cases participants noted that the software was preferable for rote practice of applying concepts, rather than learning new concepts. In this study we were motivated to produce software that could show learning beyond rote practice, independent of an instructor. Based on our experience in the previous study, we were also motivated to investigate whether and how dynamic difficulty adjustment can be used to create an engaging user experience in this study.

In the final study we created an experimental iPad application to improve reading comprehension skills by teaching the question generation strategy. We also created a control application that only allowed users to self-practice their reading comprehension skills, without teaching the reading comprehension strategy. The applications both involved having users read a passage of text, followed by answering a series of questions about the passage. The experiment iPad application adjusted the difficulty level of the passage that the user was given over time according to their performance in the application, but the control application did not. The application was tested with 48 undergraduate students at McMaster University, and we found that, based on pre- and post-session quizzes, the experimental application improved participant reading comprehension. We found that our dynamic difficulty adjustment design approach was poorly received by some participants, which may explain why there was no difference in engagement between the experiment and control applications.

Chapter 2

Scroll Shooter Study

The following work was published in the journal of Entertainment Computing:

K. Browne, and C. Anand, An empirical evaluation of user interfaces for a mobile video game, Entertainment Computing 3.1 (2012) 1-10. (doi:10.1016/j.entcom.2011.06.001)

2.1 Abstract

In this paper we empirically test the effectiveness and enjoyability of three user interfaces used to play an iPod Touch scroll shooter video game. Mobile devices are currently undergoing a surge in market penetration both in business and with consumers. These devices allow for user interface options such as touch screens and accelerometers, which are novel to mobile platforms and to large portions of the general public. To explore the effectiveness and enjoyability of these user interface options, the game was implemented with an accelerometer based interface, a touch screen based interface involving simulated buttons and a touch screen based interface involving finger gestures. The game has been formally tested with 36 human subjects each playing the game with each of the three interfaces. We present statistically significant results that the accelerometer based interface was the preferred interface and the interface in which participants performed best. We hope manufacturers will consider using the approach used in this paper to test user interfaces in-house before releasing them, since, as we show, it is inexpensive to obtain statistically significant results. We propose heuristics for mobile user interface design based on an analysis of the results and suggest an avenue for future work.

2.2 Introduction

Mobile computing devices such as smartphones, e-readers and tablet computers are currently undergoing a phenomenal surge in market penetration. Market tracker iSuppli Corp. expects smartphone shipments to rise 105% from 246.9 million in 2010 to 506 million units in 2014 [She10]. Shipments of tablet computers like the iPad are expected to grow from 19.5 million units in 2010 to 208 million units in 2014, according to Gartner Inc. media analysts [GAR10].

Some of these mobile computing devices, such as the iPhone, offer user interface capabilities beyond physical buttons, such as accelerometers and touch screens capable of recognizing the movements of multiple fingers. While strictly speaking these user interface capabilities are not completely novel, never before have they been available in mobile devices with such deep market penetration. According to estimates by Canalys, for instance, in Q4 2009 touch screen enabled smartphone sales grew by 138% year-on-year whereas the overall smartphone market grew by 41%, in fact, 55% of all smartphones shipped were touchscreen enabled [Can10].

Mobile video games are expected to ride the wave of popularity of these new devices over the next five years, as Strategy Analytics predicts that mobile gamers will increase by 57%, from 532.1 million users in 2010 to 835.7 million in 2015 [Jef10]. One can expect that maximizing user enjoyment of these games will be of critical interest to the firms developing them.

The surge in mobile device usage will likely have an impact far beyond video games, as researchers work towards using mobile devices for a broad range of tasks, from maintenance and inspection tasks in the rail industry [Dad09], to mobile payment systems [YLI+09], to health care delivery [CCS+09]. Given the broad range of potential applications and the recent history of computing technologies increasing productivity and standards of living, the potential for mobile computing to have similar benefits for society should be evident.

The promise of these new devices may remain unfulfilled if users resist or reject the device because the device or its software is not enjoyable to use, or is not intuitive and easy to learn. For instance, when it was first released the touch screen enabled Blackberry Storm was criticized by reviewers [Seg08, Cha08] for the difficult learning curve of its user interface. One user who returned the device described their dissatisfaction: "I found myself wanting to throw it in the ocean due to my frustration with its overall usability" [SS09]. We cannot capture the promise of these new devices if users return them to retailers (or throw them into the ocean!) out of frustration with their user interfaces.

Given the growing importance of mobile user interfaces to society, and our evident lack of mastery of the subject thus far, we are motivated to explore what insights into mobile user interface design can be found through user studies comparing different mobile user interfaces. The main result of this work is the derivation from the results of an experiment, design heuristics (i.e. characteristics of a good user interface) and principles for better mobile user interfaces with respect to enjoyability, ease of use, learnability and user performance. As mobile video games are at present a widely popular application genre for these devices, we decided to use a mobile video game as the basis for our work.

We developed a "scroll shooter" game for the iPod Touch, a device similar to the iPhone except for the lack of cell phone capabilities. In the scroll shooter game, the user controls a spaceship capable of firing projectiles with the aim of destroying incoming enemies, which themselves are also capable of firing projectiles. The enemies appear at the top of the screen and move downwards across the screen until they are either destroyed or "disappear" off the bottom edge of the screen (hence the game may be described as a vertical scroll shooter).

The game can be played with three different user interfaces: accelerometer, touch screen gesture ("touch gesture") and touch screen simulated button ("simulated button"). In the accelerometer interface, the user plays the game by tilting and shaking the device. Within the touch gesture interface the user plays the game by swiping their finger across the touch screen and tapping the touch screen. Finally, with the simulated button interface the user plays the game using a simulated directional pad (d-pad) and button.

The more novel work of this paper is the discussion and conclusions derived from the results of the experiment, with the game itself containing no real groundbreaking innovations. However, an overview of the game itself is presented in Section 2.3 as it is very important to give context to the results presented here.

The experiment involved having the participants play the game with each of the three interfaces, during which we measured their performance with each interface and recorded their feedback via questionnaires. Aside from the primary goal of deriving more general design heuristics and principles for mobile user interfaces, we have achieved other peripheral and related research objectives using the results gathered, such as determining which user interface would be most preferred, and which would elicit the highest in-game performance. In both cases, we found it to be the accelerometer interface. The design of the user study conducted is outlined in Section 2.4 and the results are presented in Section 2.5 with some discussion.

Much work has been done in the area of empirically studying user interfaces for video games and mobile devices. However, we could not find any work comparing our selection of user interface options (which we felt derived naturally from the capabilities of the iPod Touch) with this type of mobile video game. Given the success of the iPhone and the iPod Touch, and mobile gaming on these devices through the App Store, we found this surprising.

In work done by Gilbertson et al [GCCV08], accelerometer based control of a 3D first-person driving game called *Tunnel Run* is compared experimentally with physical button (joystick) based control. In a paper by Wei et al [WMG08], a new touch screen finger gesture based interface for playing first person shooters on a PDA device is developed and tested experimentally against a physical button interface. While not a comparative study, in work published by Chehimi et al [CC08], an accelerometer based interface is developed for a 3D space shooter game and was empirically tested with a group of participants. None of these papers use simulated button interfaces, an interface of interest given the rise in popularity of touch screen interfaces for mobile devices, and none of these papers compare our chosen selection of user interfaces.

In a paper by Rouanet et al [RBO09], three user interfaces for navigating a robot through an obstacle course are tested empirically. Two of these user interfaces are implemented using the iPhone; one is a simulated keypad and another uses touch screen gestures. The third is an accelerometer based interface implemented using the Nintendo Wii Remote. While the selection of user interfaces in this study is very similar to our selection of user interfaces, the problem domain of navigating a physical robot in real space is completely different. Our experiment design was influenced by this paper in particular, but also the mobile video game related papers cited.

2.3 Scroll Shooter

We developed the scroll shooter game for the iPod Touch using the game engine source code examples found in the book *iPhone Game Development: Developing 2D and 3D Games in Objective-C* by Paul Zirkle and Joe Hogue [ZH09] as a starting point. Before the game was used for the experiment, we had four colleagues informally test the game in an effort to prevent using an interface with obvious flaws in the experiment. Several changes were made as a result of this informal feedback and will be mentioned. A screenshot of the gameplay is shown in Figure 2.1, note that the game is exclusively played and held in the "portrait mode" of the iPod Touch shown in the screenshot.

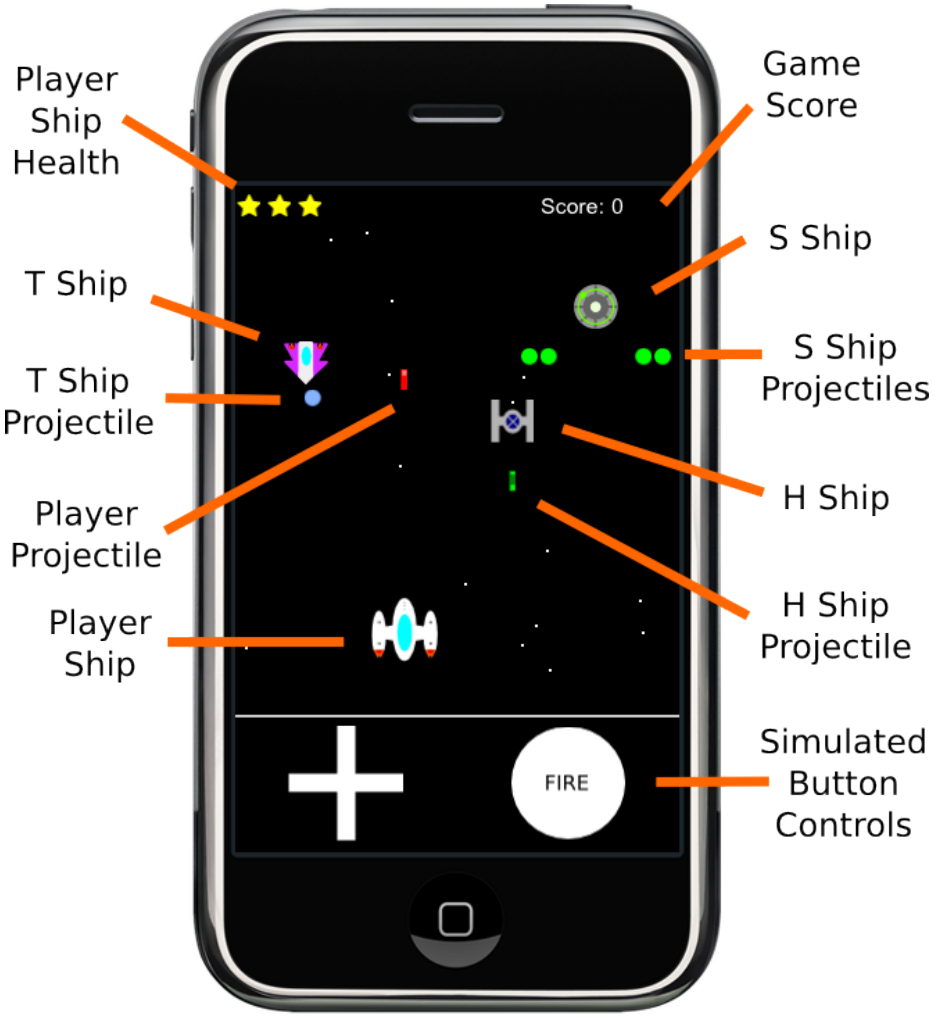

Figure 2.1: Gameplay screenshoot

The bottom portion of the screen, 100 pixels in length, is reserved for the simulated buttons in the case that the simulated button interface is selected. In the case that the touch gesture or accelerometer interface is selected, this section of the screen is left black. The player can move their ship anywhere within the top 380 x 320 pixel portion of the screen where the actual gameplay occurs. The game is played with a fixed overhead camera viewpoint, so unlike some games moving the player's ship does not cause the view of the

"'game world" itself to change.

An argument can be made that by leaving the bottom portion of the screen blank for two of the interfaces, we are forcing a limitation of the simulated buttons onto these interfaces and not presenting them in their optimal implementation. However, if the gameplay was extended into the bottom portion of the screen for the other two interfaces, the game would not be the same across all three interfaces. After some discussion we chose to keep the bottom portion of the screen blank in the interest of measuring performance with each interface without introducing an extraneous variable.

In every interface, the player's ship can move up, down, left, right, up-left, up-right, down-left and down-right. The ship moves at a constant speed in the relevant direction; there is no momentum or acceleration to the ship's movement. Finally, the player's ship can fire rectangular red projectiles, at a rate no greater than one every 185 milliseconds. If these projectiles come into contact with an enemy ship, the projectile will disappear and the enemy ship will explode and disappear.

There are three different enemy ships: H-Ships, S-Ships and T-Ships. All ships move from the top of the screen to the bottom of the screen with the same vertical velocity. T-Ships move horizontally in a sinusoidal pattern to make their destruction more difficult. S-Ships and H-Ships do not move horizontally. New ships are always initialized just outside the top of the gameplay screen, and then "fly into play". H-Ships fire rectangular green projectiles directly forward. S-Ships fire four circular green projectiles at a time, two to each side, on an angle and with slightly different velocities. T-Ships fire blue circular projectiles, which at their time of firing are given a flight path towards the current location of the player's ship on the screen, forcing the user to dodge the projectile.

The player's ship starts off with three "stars of health" as seen in Figure 2.1. If an enemy projectile or ship comes into contact with the player's ship, it is destroyed, and the player's ship loses a health star. If the player's ship loses all three health stars, it is also destroyed and gameplay concludes as a "game over" screen is presented to the user. In order to best judge the performance of the user interfaces it was decided that the users would be allotted three health stars instead of one to reduce the chances that a "fluke hit" would distort the results.

The game can be played in either a demo mode, or one of three levels. In the demo mode, H-Ships are created randomly at different horizontal positions, and fire randomly. The purpose of the demo mode is to allow users to learn how to use the different user interfaces. Each level is a stack of events which occur in sequence, with enemy ships being created at the exact same time and

position, and enemy ships firing at the exact same times, each time the same level is played. The end of a level is reached when all events have occurred and all enemy ships and projectiles are no longer in the gameplay screen, so long as the player's ship has survived.

The levels were created to be of approximately the same difficulty. The last events for each of the levels occur at 3 minutes and 30 seconds into gameplay, and as such each level can be considered to be approximately that long in length. In order to create levels of approximately equal difficulty, each level was broken down into 7 waves of 30 seconds each. Within each wave, independent of the level, the same number of T-Ships, S-Ships and H-Ships will appear, and they will fire projectiles the same number of times. However, for each level the order, position and timing in which these ships appear is different, as well as the position and timing of the projectiles. Play within each level gets progressively harder, as the number of ships created and projectiles fired in each wave increases. The details of the waves are found in Table 2.1. The levels were designed to get harder in a slow and steady fashion, to try to capture more precisely how users perform with each user interface. We believe that with this design the time played before the game over screen is reached becomes a more meaningful measure to compare user interface performance than if levels were consistently difficult, as time played reflects the ability of the player to survive at different levels of difficulty. Having the player's ship destroyed by the first wave of enemy ships should be rare for even inexperienced gamers, and beating a level should be near impossible for even the most experienced gamers.

Wave	**Time**	**Enemy Ships**	**Enemy Projectiles**
1	0-30s	5	5
2	30-60s	10	15
3	60-90s	15	25
4	90-120s	20	30
5	120-150s	25	40
6	150-180s	30	70
7	180-210s	40	165
Total:		145	350

Table 2.1: Enemy ships and projectiles for each wave

The simulated button interface works much like one would expect. Tapping the fire button fires a projectile, holding a finger or thumb down on the

fire button has no automatic fire effect, but the user can continually tap the fire button if they wish. When a finger is touching the up, down, left or right prongs of the d-pad, the player's ship will move in that direction. When a finger touches the space in between the up and right prongs of the d-pad, the player's ship will move in that direction, and so forth. The user can drag their finger from one area of the d-pad to the next, and the ship will change direction, i.e. the user does not have to lift their finger off the d-pad and retouch another area of it to change directions. The d-pad of the simulated button interface has a length of approximately 13mm, for comparison's sake the Nintendo DS Lite has a d-pad of length 18.6mm. While a larger d-pad would be ideal, it would cost more of the already limited gameplay portion of the screen. Early builds of the game used an even smaller d-pad, but the informal feedback received from colleagues for the smaller size of d-pad was very negative, whereas this size was deemed appropriate.

The touch gesture interface responds to finger swipes across the screen and taps on the screen. The tapping and finger swipe movements can occur *anywhere* on the screen, either in the top gameplay portion or in the black panel at the bottom of the screen where the simulated buttons would appear in the simulated button interface, it makes no difference. Tapping the screen fires a projectile, and again holding a finger down has no automatic fire effect, but the user can continually tap the screen if they wish. Dragging or swiping a finger across the screen will cause the player's ship to move in the up, down, left, right, up-left, up-right, down-left or down-right direction that the finger is being dragged. When the finger is lifted up off the screen, the ship will stop moving. If the finger does not lift up, but is simply held in place after dragging in a given direction, the ship will continue to move in that direction. If the finger changes direction as it is being dragged across the screen, the movement of the ship will change to reflect that new direction. So for instance in Figure 2.2, when the user started dragging their finger from point 1 to point 2, the player's ship would move in an up-right direction. When the player started moving their finger from point 2 to point 3, the ship would move in an upwards direction, until the user lifted their finger from the screen.

The accelerometer interface responds to tilting and shaking the device. When the user gently shakes the device vertically, a projectile is fired. If the user continuously shakes the device projectiles will continue to be fired, but at a rate no greater than the maximum rate of one every 185 milliseconds. Starting with the device held in portrait mode, flat and parallel to the ground, if the user tilts the left side of the device towards the ground, the player's ship moves to the left. Similarly tilting the right side of the device towards the ground moves the player's ship to the right. Tilting the top of the device

Figure 2.2: Touchscreen movement

towards the ground moves the player's ship up, and tilting the bottom of the device towards the ground moves the player's ship down. By combining tilting of the device in different directions, the ship can be moved in the standard diagonal directions.

The capabilities of the user interfaces were created to be as symmetrical as possible, so as not to give one interface any obvious advantage during gameplay. This is why in every interface only the discussed eight directions

of movement are available, the ship always has the same constant speed in whatever directions it is moving, and projectiles may only be fired at the maximum rate of one every 185 milliseconds. That said, one asymmetry is worth making note of. While it is very easy to both move the ship and fire projectiles with the simulated button and accelerometer interfaces, it is unreliable to do so with the touch gesture interface. Our attempts to use the multi-touch capabilities of the device to allow for this were unsuccessful, as tapping to fire can be interpreted by the game as a finger dragging gesture and the movement of the player's ship may be temporarily affected. During informal testing we found that solutions to allow for tapping and dragging at the same time, like setting aside a portion of the screen for taps and a portion of the screen for dragging, were rejected by our colleagues for taking away from the simplicity of the current implementation. As such, we allowed this asymmetry between the user interfaces to remain.

The game score starts at zero and increases in increments of 100 every time an enemy ship is destroyed; it is more for cosmetic effect to allow users to keep track of their progress. The game keeps track of other information during gameplay however, including:

- total time played
- number of player projectiles fired
- number of player projectiles that connected with enemy ships
- number of enemy ships destroyed
- how much time the ship spends moving in each of the eight directions
- how much time the ship spends in the nine sections of the screen denoted in Figure 2.3

We are obviously able to compute the accuracy of the user's projectile firing attempts given the total projectiles fired and the number connected, a potentially interesting statistic. The sections of the screen for which we keep track of movement are uneven in size because we found in informal testing that users simply don't move to the very top of the screen very often, for the obvious reason that it is a very risky gameplay manoeuvre.

Figure 2.3: Screen sections

2.4 Experiment Design

Experiment participants were gathered from both the student body of McMaster University and the general public. Recruitment was done with on-campus posters, department-wide e-mails to several departments in diverse areas of study, a wall post on a McMaster computer science Facebook group and broadcast-style one-to-many messages to one of the author's Facebook

friends. Word-of-mouth advertising also played a factor in recruitment as some participants recommended the experiment to others.

2.4.1 Experiment Protocol

The following protocol was followed with each experiment participant. The protocol refers to the pre-experiment questionnaire in Section 2.4.4, the interface experience questionnaire in Section 2.4.5 and the post-experiment questionnaire in Section 2.4.6.

1. The participant read and signed a consent form.
2. The participant completed a paper copy of the pre-experiment questionnaire.
3. The participant was told the goal of the experiment.
4. The rest of the experiment procedure was outlined for the participant.
5. Each user interface was demonstrated for the participant for about a minute.
6. The participant was given the chance to practice playing the game in the demo mode with each user interface (in whatever order) for no more than three minutes, or until the participant felt no more practice was necessary.
7. The participant played the game with a pre-determined user interface, on a pre-determined level.
8. The participant answered the interface experience questionnaire for that user interface.
9. Steps 7-8 were repeated for the other two user interfaces and levels.
10. The participant completed the final post-experiment questionnaire.

Because exactly 36 participants were recruited, all 36 possible combinations of level orders and interface orders were used for steps 7-8 of the protocol. Such a factorial experiment design [LSA+98] was used because if we always tested with the same level and/or interface orders, this could obviously skew the results, since, for instance, players could simply learn to play the game

better regardless of the interface by the time they get to the third time playing with a level/interface. Several test environments were used, however they all included somewhere comfortable to sit down, which all participants did. The game was played by all participants without headphones or sound. We felt it important to allow participants to acclimate themselves with the user interfaces *before* playing through the levels. We did not want participants to lose while playing through a level because they were still learning how to use the interface on a basic level.

2.4.2 Quantitative Observations

The quantitative observations were made using the data recorded by the game such as the total time played, ship movement, projectile accuracy, etc., which were discussed in Section 2.3. These quantitative observations are only recorded when the participants play through the levels, and not in the demo modes.

Observations as to how the participant perceived each user interface were recorded with the interface experience questionnaire of Section 2.4.5. The data is given numerical values in order to quantify the experiences relative to one another in more precise terms than English descriptions of the experiences could provide.

2.4.3 Qualitative Observations

The post-experiment questionnaire of Section 2.4.6 allowed for more informal English descriptions as to which user interface the participant preferred and why they preferred that interface. Any expressions that the participants made, such as expressions of frustration or jubilation, were recorded or made note of as well.

2.4.4 Pre-Experiment Questionnaire

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

- Gender (M/F/Other)
- Handedness (Right/Left/Ambidextrous)
- Student at or graduated from a Computer Science or Software Engineering University or College program (Y/N)

- Student at or graduated from any University or College program (Y/N)
- Age

The participants were also asked to rate their expertise with the following different media and interfaces: mobile phones, console or PC video games, iPhone / iPod Touch (general usage), iPhone / iPod Touch video games, touch screen interfaces and accelerometer controls (e.g. Wii). Participants rated their expertise by selecting one of the following expertise levels, based on the description of the expertise level given to them:

1. **No Expertise** I have never or almost never used this media/interface. I am not sure how to use this media/interface at all.
2. **Some Expertise** I use this media/interface 0-1 times a week on average. I can accomplish what I want using this media/interface, but do not feel sure that I know how to use it properly.
3. **Typical Expertise** I use this media/interface 2-3 times a week on average. I can accomplish what I want using this media/interface, and I feel sure that I know how to use it properly.
4. **Above Average Expertise** I use this media/interface 4-5 times a week on average. I would feel comfortable explaining how to use this media/interface to a friend.
5. **Expert** I use this media/interface 5+ times a week on average. I would feel comfortable writing an instruction manual on how to use this media/interface, including more advanced capabilities.

For analysis purposes these expertise levels were assigned the numeric values 1-5 from no expertise to expert.

2.4.5 Interface Experience Questionnaire

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

- **S1** Using this interface was enjoyable.
- **S2** Learning this interface was easy.
- **S3** This interface was comfortable to use.

- **S4** Moving the ship was easy with this interface.
- **S5** Firing at enemies was accurate with this interface.
- **S6** Evading enemies and enemy projectiles was accurate with this interface.
- **S7** My intended actions were accurately carried out on screen when using this interface.
- **S8** I feel like more practice time with this interface would have made a significant difference to my performance.

The participants could choose from: disagree, somewhat disagree, neutral, somewhat agree and agree. Again for analysis purposes these descriptions were assigned numeric values 1-5 from disagree to agree.

2.4.6 Post-Experiment Questionnaire

The following questions were asked on the post-experiment questionnaire. Beyond the restriction of selecting only one most preferred and only one least preferred user interface, this section was intended to be relatively "free form" where subjective experiences could be surveyed in a non-numerical manner. As such, participants were given boxes to write down why an interface was their most preferred and why an interface was their least preferred.

1. Which interface was your most preferred user interface? Select only one.
2. Why was this your most most preferred user interface?
3. Which interface was your least preferred user interface? Select only one.
4. Why was this your least preferred interface?
5. Is there anyway that you believe this experiment's participant experience could have been improved?

2.5 Results and Discussion

The 36 experiment participants, 13 females and 23 males, ranged in age from 19 to 43 with an average age of 26 and a standard deviation of 4.36. There were 4 left-handed, 2 ambidextrous and 30 right-handed participants. There were

26 participants who were either a student at or who had graduated from either a University or College program, and 13 of those graduated from a Computer Science or Software Engineering program.

With a sample group of this age and education range, the results of this study cannot be extended in a statistically significant way to the general population of the world. However, given the reasonably random participant selection process, we believe some of our results are statistically significant for the sampled population of those who came in contact with the recruitment procedures that were outlined in Section 2.4.

When we talk about results being statistically significant for the population, it is this population we refer to and not the general population. If we cannot say something about a result with a p-value of 0.05 or less, we do not consider the result to be statistically significant. In the case of establishing statistical significance for proportions, we make inferences based on the z-score for a sample proportion. In the case of establishing statistical significance for mean values, we compute the z-score for that sampled mean value. When we discuss one mean value's significance relative to another, we compute the difference of the values in question for each participant, and then compute the proportion of participants for which a value is greater. We can then make inferences based on the z-score for this proportion.

The data regarding participant expertise with various media and interfaces which was collected in the pre-experiment questionnaire is presented in Table 2.2. The average reported expertise with most of the interfaces/media was between "below average expertise" and "average expertise".

Interface/Media	**Avg.**	**SD**
Mobile phones	3.58	1.11
Console or PC video games	3.28	1.11
Mobile video games	2.64	1.1
iPhone / iPod Touch (general usage)	2.44	1.42
iPhone / iPod Touch (video games)	2.19	1.33
Touch screen interfaces	2.72	1.11
Accelerometer controls (e.g. Wii)	2.64	1.13

Table 2.2: Participant expertise data - average and standard deviation

The percentage of participants who most preferred each user interface is presented in Figure 2.4. One very statistically significant result was the

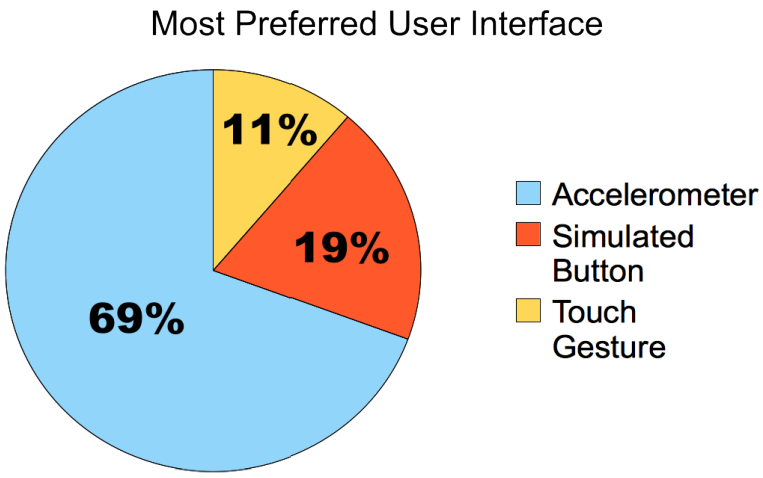

Figure 2.4: Most preferred user interface

overwhelming preference for the accelerometer interface. We can say with statistical significance (p-value 0.02) that the majority of the population most prefers the accelerometer interface. We cannot make a statistically significant comparison between the simulated button and touch gesture user interfaces.

The percentage of participants who least preferred each user interface is presented in Figure 2.5. Unfortunately given how close the percentages are we cannot say anything statistically significant about which interface is least preferred from this data.

In the written feedback the participants gave several common reasons describing why the accelerometer interface was their most preferred. Virtually all of them wrote that the controls were either "intuitive", "natural" or "easy to learn". Other common reasons cited included the fact that it was easiest to fire projectiles and move at the same time, that their hands weren't covering the screen and that the interface was more fun because this interface was a new type of experience for them. Amongst the comments of those participants who most preferred either the simulated button or touch gesture interfaces there was no real common thread as to why they most preferred those interfaces.

The written feedback describing why each user interface was least preferred was fairly informative. In the case of the touch gesture interface, many participants complained that it was too difficult to fire and move at the same time. Several participants complained that their fingers would block the screen while playing the game. Several participants complained about the sensitivity

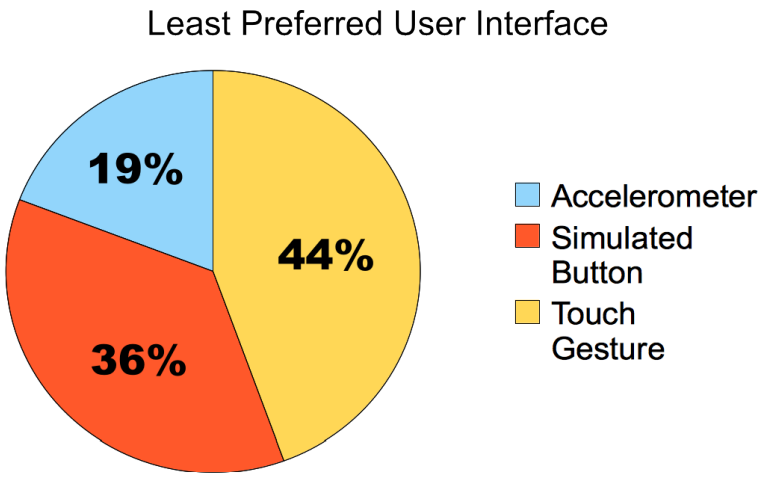

Figure 2.5: Least preferred user interface

of the ship's movement relative to their finger's movement, one describing it as like a "low sensitivity mouse". Talking to the participants, we believe what they expected was for the ship to move either *as fast* as they swiped their finger and/or to *where* they swiped their finger. While this sort of implementation was not possible for this experiment in the interest of keeping the capabilities of the user interfaces symmetrical, we believe the participants make an excellent point. We believe either of these suggestions would give the interface a more intuitive feel as more properties of the finger's physical movement, such as speed or location, would be directly translated into the properties of the ship's movement in the game. A final point of particular interest with this user interface was that one female participant with long finger nails could not effectively control the ship with this interface, as both the skin and nail of her finger were contacting the device, and as a result this interface was essentially useless for her.

In the case of those who least preferred the simulated button interface, a majority complained about a lack of responsiveness or a lack of accuracy in regards to ship movement. We do not believe this is due to a lag on a part of the game in detecting touches on the d-pad. We believe after talking to participants what was happening is that nearly the entire thumb of the participant was covering the d-pad, and so with nothing to see or feel, they were moving their finger around like it was a joystick or analog stick. However, if for instance a user who was moving right decided to move left, there would

be a delay in the time that they started dragging their finger left and the time their finger was over the left prong of the d-pad, causing a perception of delayed response and/or inaccuracy. Finally a few others complained that the simulated button interface was uncomfortable, or that their fingers felt crowded on the screen.

There were two common complaints amongst those who least preferred the accelerometer interface. One was that focusing on the screen was difficult when the screen was being moved in order to play the game; a couple of participants complained about screen reflections in particular. The other was that shaking to fire was imprecise compared to tapping the screen or a simulated button.

Table 2.3 contains the population's average response to the interface experience questionnaire statements (found in Section 2.4.5), computed from our sample results with 95% confidence interval. The results for S1 indicated with statistical significance (p-value 0.05) that the majority of the population would more strongly agree that the accelerometer interface is enjoyable than either of the two other interfaces.

	Touch Gesture	**Accelerometer**	**Simulated Button**
S1	3.69 ± 0.32	4.50 ± 0.23	3.22 ± 0.36
S2	4.36 ± 0.29	4.58 ± 0.23	4.47 ± 0.32
S3	3.72 ± 0.33	3.97 ± 0.33	3.11 ± 0.42
S4	3.42 ± 0.42	4.14 ± 0.35	3.00 ± 0.49
S5	3.56 ± 0.38	3.61 ± 0.36	4.08 ± 0.35
S6	3.19 ± 0.38	3.94 ± 0.32	3.08 ± 0.39
S7	3.11 ± 0.37	4.11 ± 0.29	3.14 ± 0.45
S8	3.65 ± 0.38	4.36 ± 0.26	3.97 ± 0.35

Table 2.3: Interface experience questionnaire population average (95% confidence interval), computed from sample results

Looking at the results for S2 we should note that the percentage of the participants that either somewhat agreed or agreed that an interface was easy to learn was 81% for the touch screen gesture interface, 97% for the accelerometer interface and 92% for the simulated button interface. As a result we can say with statistical significance (p-value 0.01) that the majority of the population would either somewhat agree or agree that the interfaces

were easy to learn. We believe this adds some credibility to our experiment method and results, as the vast majority of participants were not struggling to learn the interfaces themselves. The higher mean values of the accelerometer interface for the remainder of the statements is consistent with the preference for the accelerometer interface, and the interface performance statistics to be discussed.

The results with regard to user interface performance contained some statistically significant differences. Likely the best metric of how "well" participants performed in the game with each user interface overall was how long they lasted playing through a level before having their ship destroyed (no player completed/"beat" a level, as intended). Figure 2.6 shows the average time at which game over occurred with each user interface.

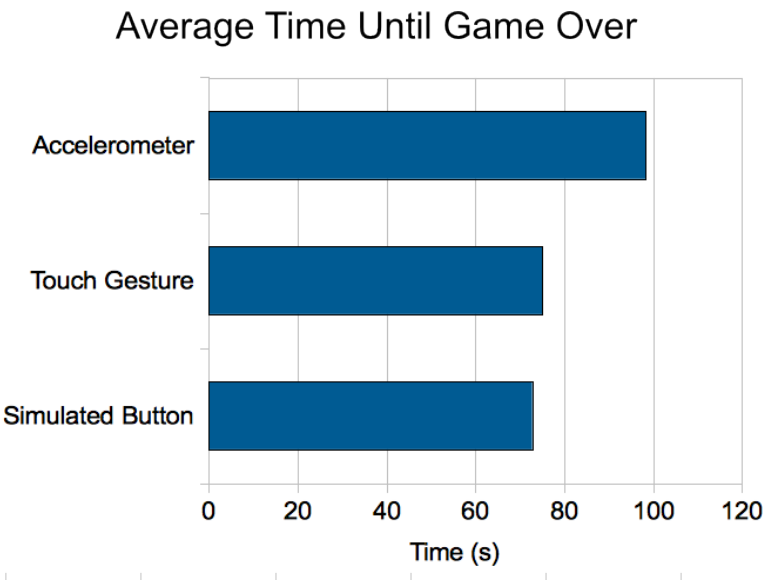

Figure 2.6: Average time game over was reached for each user interface

As participants played through three levels each, one for each interface, we observed that 67% lasted the longest with the accelerometer interface, 19% with the simulated button interface and 14% with the touch gesture interface. We can therefore say with statistical significance (p-value 0.04) that the majority of the population would persist longer with the accelerometer interface than the other interfaces. On the other hand, 11% lasted the smallest amount of time in a level with the accelerometer interface, 42% with the simulated button interface and 47% with the touch gesture interface. However, these differences are not statistically significanct.

The gameplay statistics relating to the firing of player projectiles are presented in Table 2.4. Some context is needed to interpret these results properly; the insight provided by the total shots fired and total shots connected statistics is limited by the fact that on average participants played much longer with the accelerometer interface than the others. As such we instead look at shots fired per second of gameplay, and we find with statistical significance (p-value 0.001) that the majority of the population fires more shots per second with the accelerometer interface than the touch gesture interface, but are unable to make further statistically meaningful comparisons. The shot accuracy statistics show no statistically significant differences between interfaces, which is consistent with the perception of participants reflected in the interface experience questionnaires.

Touch Gesture				
	Avg.	**Min**	**Max**	**SD**
Total Shots Fired	80.14	17	316	73.62
Total Shots Connected	18.86	2	63	73.62
Shot Accuracy	28%	7%	94%	16%
Shots Fired Per Second	0.99	0.43	3.36	0.64
Accelerometer				
	Avg.	**Min**	**Max**	**SD**
Total Shots Fired	150.53	34	431	92.15
Total Shots Connected	32.61	8	91	18.53
Shot Accuracy	24%	5%	42%	10%
Shots Fired Per Second	1.49	0.49	3.57	0.71
Simulated Button				
	Avg.	**Min**	**Max**	**SD**
Total Shots Fired	117.06	2	564	40.37
Total Shots Connected	21.19	1	107	21.80
Shot Accuracy	23%	7%	53%	12%
Shots Fired Per Second	1.33	0.18	3.7	0.89

Table 2.4: Projectile-related gameplay statistics

Performance differences between the user interfaces in regards to movement of the player's ship also showed some statistically significant differences. We recorded how long participants moved the player's ship in each of the eight

available directions, and this did not directly give us any interesting data because given the "boxed" nature of the gameplay a ship moves left about as much as it does right and down about as much as it does up. However, as presented in Table 2.5, we also looked at the percentage of time that a ship moved during gameplay (PTM) and the percentage of *that* time that the ship moved along one of the four available diagonal directions (PD). We found no statistically significant difference between the user interfaces in regards to PTM, however we found with statistical significance (p-value 0.001) that the majority of the population will move more diagonally with the touch gesture user interface than the other user interfaces.

Touch Gesture				
	Avg.	**Min**	**Max**	**SD**
PTM	43%	19%	76%	14%
PD	31%	7%	69%	15%
Accelerometer				
	Avg.	**Min**	**Max**	**SD**
PTM	44%	21%	77%	12%
PD	7%	0%	31%	7%
Simulated Button				
	Avg.	**Min**	**Max**	**SD**
PTM	50%	25%	73%	12%
PD	11%	0%	35%	8%

Table 2.5: Movement-related gameplay statistics

Recalling the gameplay screen positions of Figure 2.3, the percentage of time (95% confidence interval) that the population will have the player's ship located in each position is presented in Table 2.6. One thing that we noticed observing the participants playing the game was that they tended to stay at the bottom of the screen (positions 1-3) during much of the gameplay, a good strategy to be sure, but that they did so less in the case of the touch gesture interface. Strictly looking at the average percentage of time spent at any of the positions 4-9, the portion of the gameplay screen above positions 1-3, we found values of 44.10%, 19.43% and 29.48% for the touch gesture, accelerometer and simulated button interfaces respectively. We are able to say with statistical significance (p-value 0.001) that the majority of the population will spend more

time in positions 4-9 with the touch gesture interface than the accelerometer interface, and with statistical significance (p-value 0.01) that the majority of the population will spend more time in positions 4-9 with the touch gesture interface than with the simulated button interface.

Pos.	**Touch Gesture**	**Accelerometer**	**Simulated Button**
1	10.6 ± 2.3	18.2 ± 3.1	19.5 ± 3.6
2	33.6 ± 5.6	40.7 ± 4.2	35.0 ± 3.6
3	11.7 ± 2.4	21.7 ± 3.2	16.1 ± 2.7
4	5.5 ± 1.6	3.1 ± 1.2	4.5 ± 1.0
5	21.6 ± 4.3	8.5 ± 3.7	8.3 ± 2.1
6	5.6 ± 1.5	3.0 ± 1.1	4.1 ± 1.4
7	1.5 ± 0.7	1.2 ± 0.9	4.1 ± 1.9
8	8.1 ± 3.4	2.7 ± 1.5	6.4 ± 2.9
9	1.8 ± 0.8	1.0 ± 0.53	2.1 ± 0.8

Table 2.6: Player's ship position population average (95% confidence interval), computed from sample results

We have found that the touch gesture interface led to more diagonal movement and for the player's ship to be more likely positioned at higher positions of the gameplay screen than other interfaces. The interface experience questionnaire response for S7 found in Table 2.3 does not show strong disagreement with regards to whether participants felt their intended actions were carried out on screen correctly using the touch gesture interface, indicating that the actions the users did carry out were intended. This suggests that the differences in gameplay style observed for the touch screen interface were really intended by the participants and not accidental.

Looking at how closely aligned preference for a user interface was aligned with performance within a user interface, all but 5 participants most preferred the user interface that they performed best in the game with (lasted longest until game over), and all but 11 participants least preferred the interface that they performed worst with. It is possible that preference for an interface causes better performance with that interface, or that better performance with an interface causes a preference for that interface. Determining if either of these situations was the case for our experiment was beyond the aims of this research, but we note that the literature indicates that preference for and performance with an interface are not necessarily correlated [ACD95].

We also tried to find any statistically significant correlations between gender, age or prior experience with media/interfaces and either performance with or preference for user interfaces. One interesting gender difference was that every single participant who most preferred the simulated button interface was male, and 8 out of the 13 females least preferred the simulated button interface. We can speculate that perhaps this is due to the fact that more males grew up playing video game consoles which use d-pads, however the pre-experiment questionnaire does not show a statistically significant difference between male and female expertise with console and PC video games. Other than this however, any differences found when comparing results that have been filtered for certain participant characteristics did not yield statistically meaningful or noteworthy results.

2.6 Conclusion

The main purpose of this work was to derive design heuristics for mobile user interfaces based on the data and insights from our experiment. We must be cautious and note that our design heuristics are based on a single experiment involving one type of mobile video game, with a sample not large or diverse enough to be representative of the human population as a whole. In particular, we believe our design heuristics for mobile user interfaces are relevant for mobile video games and non-games with input requirement scenarios that may resemble the gameplay of the scroll shooter game. That said, we believe that this study shows that it is relatively inexpensive to get statistically significant results regarding user interface preferences. We hope that manufacturers will consider using the approach used in this paper to test user interfaces in-house before releasing them.

Our list of design heuristics in no particular order of significance is as follows:

- **H1** An accelerometer-based user interface should be available.
- **H2** Multiple user interfaces should be available.
- **H3** Touch gestures should be utilized when diagonal direction input is either required from or desired by the user.
- **H4** Interface sensitivity should be configurable.
- **H5** Physical properties of gestures should be directly translated into virtual properties.

Firstly, regarding heuristic H1, we believe that considering the strong consensus found in our results that the accelerometer interface was both most preferred by the participants and elicited the best participant performance, that an accelerometer-based interface should essentially always be made available if feasible. Though they were the minority, given the significant number of participants who most preferred other interfaces or performed better using other interfaces we suggest in heuristic H2 that a multiple user interface implementation of a game is the best solution if feasible. We also suggest this sort of multiple interface implementation because of the participant who was simply unable to use an interface effectively due to her finger nail length; it may be possible that some users just *cannot* use certain interfaces.

Looking at heuristic H3, we believe that for instances where diagonal direction input is either required from or desired by the user, an interface similar to our touch gesture interface is preferred. This is based on our results showing that participants moved diagonally more with the touch gesture interface than the others. We believe users were simply finding it easier to essentially draw a diagonal line with their finger across the screen to move diagonally than to alternatively try to dip the device in two different directions at the same time to move diagonally.

Regarding heuristic H4, the written and verbal feedback provided to us by the participants indicated that they wished they could manipulate the sensitivity of the touch gesture user interface. Much like an operating system allows one to manipulate the sensitivity of a mouse, we believe the same should be true of touch screen and accelerometer interfaces which depend upon physical gestures.

Finally, H5 is the heuristic we believe may prove to be the strongest and most generalizable. We strongly believe based on the written and verbal user feedback that disappointment with the touch gesture interface was related to the fact that gestures on the touch screen did not lead to "intuitively expected" results. Participants would swipe their finger up and to the right and their ship would move up and to the right, but participants would complain that it would not move as fast as they swiped their finger or to the position that they swiped their finger. We believe what was actually at issue was that the participants felt intuitively that the ship's movements would mimic the real-world properties of their gesture movement. When these physical properties were not translated into the gameplay, that intuition was left unrealized and the participant had to "re-think" what they knew about movement for the game world.

We speculate that this was not the case for the accelerometer interface because the participant's intuition about the physical properties of the gestures

involved was more directly translated into the game world. If a marble were resting on the iPod Touch as it was being held flat and parallel to the ground, and a user were to tilt the device downward in a given direction, the marble would move (and ultimately fall off) in that direction. Similarly, with our accelerometer interface the player's ship moves in the direction that the user tilts the device. Physical properties and intuitions about movement in this case are translated into the gameplay.

Heuristic H5 essentially works under the premise that we should not force users to (re)learn anything. They should be able to use their pre-existing understanding of the physical world to control the device as much as possible. We suggest that in many problem domains this is best accomplished when as many of the physical properties of the gesture as possible are translated as directly as possible into the virtual properties of the interface's reaction to that gesture.

In order to determine how novel this idea is we have conducted a literature search. We believe this idea is similar to the direct manipulation human-computer interaction style [Shn84], where virtual interfaces are designed using an appropriate physical model of reality, in the sense that the user's understanding of the physical world is used to help them more easily use a virtual interface. This idea is also related to Affordances, situations where an object's sensory characteristics intuitively imply its functionality and use, as proposed by Donald Norman [Nor88]. The *principle of the moving part* of the thirteen principles of display design advocated by Wickens et al [WLLGB03] suggests that moving elements on a display which represents some real world system move in a way which is compatible with the user's mental model of that system. Again, this is an example of an understanding of the physical world being used to help the user more intuitively understand the interface.

Turning our attention towards industry documentation for mobile user interfaces, we looked at the user interface guidelines provided by Research in Motion for the Blackberry devices [Res10], the iPhone human interface guidelines provided by Apple [App10], the human interface guidelines provided by Palm for webOS [Pal10] and finally the user interface guidelines for Android provided by Google [Goo10]. All of these documents contain the user interface guidelines (i.e. principles and heuristics) currently being suggested to the developers working with these platforms. Apple's human interface guidelines suggest that a developer should, when possible, "model your application's objects and actions on objects and actions in the real world", which is related to our suggestion but not the same. Though in our literature search we found arguments for incorporating understanding of the physical world in gestures, we couldn't find anything advocating specifically what we are advocating with

this heuristic regarding physical property translation.

As an example of what we are suggesting, consider the swiping motion a user makes with their finger to go from one page of applications to the next on an iOS device such as an iPod Touch. This could have been programmed in such a manner that any swipe from right to left on the screen would instantly move the screen a page over to the right, which would still take advantage of the user's understanding of the physical world. But instead the page *also* flips over at a speed relative to how quickly you swiped your finger. Strictly speaking this is not necessary, as the device would be as capable of switching pages without this feature of the interface, but we believe this gives the device a more natural feel and makes it more enjoyable to use. What we suggest is not just that the expected result of a gesture in the physical world be reproduced in the interface, but that as many physical properties as possible of the gesture are translated into the effect of the gesture.

However, we caution that this approach may be less useful in problem domains in which translating as many physical world properties of the gesture as possible into the state of the interface makes the software *too hard* to use. For example, we would expect that a golf game with an accelerometer interface that fully translated every aspect of a real world golf swing into a virtual world would become *less* satisfying for some participants who lack the ability to swing a golf club properly. We have no data to suggest that this is the case, but it is something we feel should be explored.

We propose for future work further experiments which would investigate turning heuristic H5 into a stronger mobile user interface design principle:

Principle of physical property translation for gestures

- Whenever possible the interface should emulate behaviour expected in the physical world in response to gesture interactions.

- As many of the physical properties (speed, position, etc.) of touch screen and accelerometer gestures as possible should be translated as directly as possible into virtual properties of the in-game (or in-software) change in state, so long as these properties are reasonably within the abilities of the user.

We propose this principle of physical property translation for gestures as a hypothesis which should be tested with experiments designed to quantify it in a diverse set of problem domains, ideally with a much larger and more diverse set of participants. For a diverse set of problem domains, different

gesture-based interfaces which translate progressively more physical properties of the gesture into the gameplay or software state should be implemented, and then tested against the set of participants for performance and preference differences. If the principle is not falsified by these experiments then we believe the next step would be the construction of formalized laws quantifying expected user satisfaction for gesture-based interfaces which take into account user expertise and ability as well as the number and type of physical properties being translated, and the fidelity of those translations.

Regarding generalizability of the heuristics beyond mobile video games with similar input requirements, we believe that heuristics H1 and H3 may be less generalizable than the other heuristics. For example, if multiple user controlled objects are on screen, heuristic H1 may simply not apply as the input requirements are different than the scroll shooter. Regarding heuristic H3, it is possible that waving the device could provide diagonal direction input as effectively as the touch screen gesture interface, but our experiment did not cover this possibility. We believe that heuristics H2, H3 and H5 have the most potential to be generalized but believe that such generalization is dependent on conducting further experiments.

2.7 Bibliography

[ACD95] A. D. Andre and Wickens C. D. When users want what's not best for them. *Ergonomics in Design*, 3(4):10–14, October 1995.

[App10] Apple Inc. iPhone Human Interface Guidelines, August 2010. 2010-08-03.

[Can10] Canalys. Majority of smart phones now have touch screens. `http://www.canalys.com/pr/2010/r2010021.html`, February 2010. Last accessed August 31st, 2010.

[CC08] Fadi Chehimi and Paul Coulton. Motion controlled mobile 3d multiplayer gaming. In *ACE '08: Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology*, pages 267–270, New York, NY, USA, 2008. ACM.

[CCS+09] Sutirtha Chatterjee, Suranjan Chakraborty, Saonee Sarker, Suprateek Sarker, and Francis Y. Lau. Examining the success factors for mobile work in healthcare: A deductive study. *Decis. Support Syst.*, 46(3):620–633, 2009.

[Cha08] Bonnie Cha. RIM BlackBerry Storm (Verizon Wireless). `http://reviews.cnet.com/smartphones/rim-blackberry-storm-verizon/4505-6452\_7-33311850.html\#reviewPage1`, November 2008. Last accessed August 31st, 2010.

[Dad09] Yasamin Dadashi. *Fundamental Understanding and Future Guidance for Handheld Computers in the Rail Industry*. PhD thesis, University of Nottingham, 2009.

[GAR10] Gartner says worldwide media tablet sales on pace to reach 19.5 million units in 2010. `http://www.gartner.com/it/page.jsp?id=1452614`, October 2010. Last accessed May 28th, 2011.

[GCCV08] Paul Gilbertson, Paul Coulton, Fadi Chehimi, and Tamas Vajk. Using "tilt" as an interface to control "no-button" 3-d mobile games. *Comput. Entertain.*, 6(3):1–13, 2008.

[Goo10] Google Inc. User Interface Guidelines. `http://developer.android.com/guide/practices/ui_guidelines/index.html`, August 2010. Last accessed September 2nd, 2010.

[Jef10] Judy Jefferson. Mobile Gaming Expected To Soar as Focus Turns to Cell Phones. `http://www.brighthand.com/default.asp?newsID=16916\&news=Mobile+Gaming+Consoles+Decline`, August 2010. Last accessed August 31st, 2010.

[LSA+98] Torbjrn Lundstedt, Elisabeth Seifert, Lisbeth Abramo, Bernt Thelin, sa Nystrm, Jarle Pettersen, and Rolf Bergman. Experimental design and optimization. *Chemometrics and Intelligent Laboratory Systems*, 42(1-2):3 – 40, 1998.

[Nor88] Donald A. Norman. *The Psychology Of Everyday Things*. Basic Books, June 1988.

[Pal10] Palm Inc. Human Interface Guidelines. `http://developer.palm.com/index.php?option=com_content&view=article&id=1836&Itemid=52&limitstart=2`, August 2010. Last accessed September 2nd, 2010.

[RBO09] Pierre Rouanet, Jerome Bechu, and Pierre-Yves Oudeyer. A comparison of three interfaces using handheld devices to intuitively drive and show objects to a social robot: the impact of underlying metaphors, 2009.

[Res10] Research In Motion. Blackberry Smartphones UI Guidelines, April 2010. Version 2.4.

[Seg08] Sascha Segan. RIM BlackBerry Storm 9530 (Verizon). `http://www.pcmag.com/article2/0,2817,2331977,00.asp`, November 2008. Last accessed August 31st, 2010.

[She10] Ian Sherr. Global smartphone shipments to double in four years - iSuppli. `http://www.totaltele.com/view.aspx?ID=456081`, June 2010. Last accessed August 31st, 2010.

[Shn84] Ben Shneiderman. The future of interactive systems and the emergence of direct manipulation. In *Proc. of the NYU symposium on user interfaces on Human factors and interactive computer systems*, pages 1–28, Norwood, NJ, USA, 1984. Ablex Publishing Corp.

[SS09] Amol Sharma and Sara Silver. BlackBerry Storm Is Off To Bit of a Bumpy Start. `http://online.wsj.com/article/NA\_WSJ\_PUB:SB123292905716613927.html`, January 2009. Last accessed August 31st, 2010.

[WLLGB03] Christopher D. Wickens, John Lee, Yili D. Liu, and Sallie Gordon-Becker. *Introduction to Human Factors Engineering (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003.

[WMG08] Chen Wei, Gary Marsden, and James Gain. Novel interface for first person shooting games on pdas. In *OZCHI '08: Proceedings of the 20th Australasian Conference on Computer-Human Interaction*, pages 113–121, New York, NY, USA, 2008. ACM.

[YLI+09] Tetsuo Yamabe, Vili Lehdonvirta, Hitoshi Ito, Hayuru Soma, Hiroaki Kimura, and Tatsuo Nakajima. Applying pervasive technologies to create economic incentives that alter consumer behavior. In *Ubicomp '09: Proceedings of the 11th international conference on Ubiquitous computing*, pages 175–184, New York, NY, USA, 2009. ACM.

[ZH09] Paul Zirkle and Joe Hogue. *iPhone Game Development: Developing 2D and 3D Games in Objective-C*. O'Reilly Media, 2009.

2.7 Amendments

In this section we provide clarification, corrections and additional analysis based on feedback provided by a reviewer.

In Section 2.7.1 we discuss how the confidence intervals presented in this paper were calculated, and present additional calculations. In Section 2.7.2 we make corrections to the usage of the term "population".

In Section 2.7.3 we clarify our usage of the term "statistical significance". In Section 2.7.4 we present the z-test hypothesis tests done in this paper in detail, and in Section 2.7.5 we present binomial distribution hypothesis tests.

2.7.1 Confidence intervals

In Table 2.3 the average participant responses to the interface experience questionnaire statements are presented, along with a 95% confidence interval. We did not specify how these confidence intervals were calculated.

The confidence intervals presented in Table 2.3 were calculated using a z-score, as in the following formula:

$$x \pm z_{\alpha/2} * (s/\sqrt{n})$$

where x is the sample mean, s is the sample standard deviation, n is the sample size, $\alpha = 1 - (95/100)$ (95 being the confidence level), and z the normal distribution z-score.

The z-score was originally used to calculate the confidence intervals based on the guidelines presented in various statistics learning materials that due to the central limit theorem, for "sufficiently large" sample sizes ($n \geq 30$), the z-score may be used.

Alternatively, the t-distribution changes shape as a function of the sample size, providing a different sampling distribution for each sample size. The larger the sample size, the closer the t-distribution resembles the z-distribution. But until the sample size approaches infinity, it is better to use a t-distribution.

As a result, we re-calculate the confidence intervals for the average participant responses to the interface experience questionnaire statements, and present the results in Table 2.7. The confidence intervals presented in Table 2.7 were calculated using a t-score, as in the following formula:

$$x \pm t_{\alpha/2} * (s/\sqrt{n})$$

where x is the sample mean, s is the sample standard deviation, n is the sample size, $\alpha = 1 - (95/100)$ (95 being the confidence level), and t the t-distribution (t-score).

	Touch Gesture	**Accelerometer**	**Simulated Button**
S1	3.69 ± 0.33	4.50 ± 0.24	3.22 ± 0.37
S2	4.36 ± 0.30	4.58 ± 0.23	4.47 ± 0.33
S3	3.72 ± 0.34	3.97 ± 0.34	3.11 ± 0.44
S4	3.42 ± 0.44	4.14 ± 0.36	3.00 ± 0.51
S5	3.56 ± 0.39	3.61 ± 0.37	4.08 ± 0.37
S6	3.19 ± 0.39	3.94 ± 0.33	3.08 ± 0.41
S7	3.11 ± 0.38	4.11 ± 0.30	3.14 ± 0.47
S8	3.65 ± 0.39	4.36 ± 0.27	3.97 ± 0.36

Table 2.7: Interface experience questionnaire results (95% confidence interval)

Similarly, in Table 2.6 we presented the average percentage of time that the participants had their player ship in each gameplay screen position, along with a 95% confidence interval. The confidence intervals in this table were also calculated using a z-score. In the same manner, we re-calculate these confidence intervals using a t-score, and present the results in Table 2.8.

Pos.	**Touch Gesture**	**Accelerometer**	**Simulated Button**
1	10.6 ± 2.4	18.2 ± 3.2	19.5 ± 3.7
2	33.6 ± 5.8	40.7 ± 4.3	35.0 ± 3.7
3	11.7 ± 2.5	21.7 ± 3.3	16.1 ± 2.8
4	5.5 ± 1.7	3.1 ± 1.3	4.5 ± 1.0
5	21.6 ± 4.4	8.5 ± 3.8	8.3 ± 2.2
6	5.6 ± 1.6	3.0 ± 1.2	4.1 ± 1.5
7	1.5 ± 0.7	1.2 ± 0.9	4.1 ± 1.9
8	8.1 ± 3.5	2.7 ± 1.6	6.4 ± 3.0
9	1.8 ± 0.8	1.0 ± 0.5	2.1 ± 0.8

Table 2.8: Player's ship position average percentage of time results (95% confidence interval)

2.7.2 Usage of the term population

In several instances we used the word "population" when we should have used "sample" or described what is happening in a different way.

On page 37 we stated that "Table 2.3 contains the population's average response" when in fact the table contains the average response for the sample. The confidence intervals presented in the table serve as interval estimates for the population average response. Similarly on page 41 in the caption of Table 2.6 we refer to the player ship position population average when the table contains the average response for the sample.

2.7.3 Usage of the term statistical significance

The term *statistical significance* was used repeatedly in the paper when discussing the results, for example on page 35: "We can say with statistical significance (p-value 0.02) that the majority of the population most prefers the accelerometer interface". We used the term statistical significance to denote when, as part of a hypothesis test, a p-value was less than a significance level (i.e. the null hypothesis was rejected).

This usage of the term was found to be unclear by a reviewer (e.g. what are the comparison groups, what is the measure statistically different from).

In Section 2.7.4, we present each individual hypothesis test conducted as part of this work in full detail.

In Section 2.7.5, we present additional hypothesis tests conducted using the binomial distribution. We have conducted this analysis for the same reason we re-calculated confidence intervals in Section 2.7.1; the z-test was used in the original hypothesis tests and it uses the normal distribution, an approximation of the binomial distribution.

Another point that is important to clarify is that the hypotheses that were tested were suggested by the data collected. If we test a hypothesis on the same data that suggested the hypothesis, we have not produced strong evidence that the hypothesis is correct. This is because if we collect enough data during an experiment, it is likely that some data can be found to support *some* hypothesis. And testing a hypothesis on the same data that suggests the hypothesis involves circular reasoning (something seems true, therefore we hypothesize that is true, we test it on the same data set, and find it is true). This can lead to false positives (type I errors).

As a result, when we use hypothesis testing in this paper we are not testing hypotheses that were developed before the data was collected. Though we do not list any hypotheses in our introduction or experiment design sec-

tions, and our experiment design is exploratory, by using the term 'statistical significance' when discussing the results we could mislead the reader into believing that we have tested hypothesis developed before the data was collected.

We had no intention to mislead readers of the paper. The inaccuracy resulted from an inaccurate understanding and misuse of the statistical tools and terminology that were used.

We should have explicitly stated or labeled that these claims of statistical significance were the results of post-hoc analysis to properly describe what was done, so that readers were not mislead.

We also note that the results of the hypothesis testing are not without value. These hypothesis tests should be viewed as post hoc analysis to assist in *formulating hypothesis* for subsequent empirical verification in future studies.

2.7.4 Z-test hypothesis test details

In this paper we used the z-test to make claims regarding the statistical significance of the results. The usage of the z-test was based on a belief that for $n = 36$, the z-test would be sufficient.

All hypothesis testing was done by analyzing proportions, rather than mean values. For example, if we wished to compare the number of shots fired per second when participants used one interface relative to another, we did not do any hypothesis testing using the mean values and standard deviations. Instead, for each participant, we would subtract one value from the other value, giving us a *proportion* of participants for which one value is greater than the other. So if a participant fired an average of 1.2 shots per second with the accelerometer interface and the same participant fired 0.9 shots per second with the simulated button interface, we would count that participant as a participant who shot more shots per second using the accelerometer interface. By looking at all participants in this way, we could come up with a proportion to anaylze with hypothesis testing.

We re-evaluate the statistical analysis and claims made in this work below, and show all the work done in each hypothesis test.

Accelerometer interface preference

On page 35 we stated that we could say with statistical significance (p-value 0.02) that the majority of the population most prefers the accelerometer interface.

Our survey indicated that 25/36 participants (69.44%) preferred the accelerometer interface. We conduct the following hypothesis test at significance

level $\alpha = 0.02$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.3267 <= z <= 2.3267$.

Note: the chosen significance level implies a critical value of 2.3267.

Computing the test statistic:

$$z = (0.6944 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 2.33$$

Applying the decision rule: $z = 2.33 > 2.3267$, therefore we reject the null hypothesis.

Statement S1 results

On page 37 we stated that we could say with statistical significance (p-value 0.05) that the majority of the population would more strongly agree that the accelerometer interface is enjoyable than either of the two other interfaces.

The survey statement "Using this interface was enjoyable" was answered for each of the three interfaces by each participant using a 5 point likert-scale (1-5). We observed that 27 out of 36 participants "more strongly agreed" with this survey statement in the case of the accelerometer interface compared to the simulated button interface. In other words, the participants ranked their agreement with this statement for each interface using the likert scale, and in the case of 27 out of 36 participants, the participant gave a response at least 1 point higher on the likert scale when comparing the responses for the accelerometer interface to those of the simulated button interface. If the likert scale is measuring agreement, then a higher score is taken to indicate higher agreement.

Therefore our survey indicated that 27/36 participants (75.00%) more strongly agreed that the accelerometer interface is enjoyable than the simulated button interface. We conduct the following hypothesis test at significance level $\alpha = 0.05$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-1.96 <= z <= 1.96$.
Note: the chosen significance level implies a critical value of 1.96.
Computing the test statistic:

$$z = (0.75 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 3.00$$

Applying the decision rule: $z = 3.00 > 1.96$, therefore we reject the null hypothesis.

We observed that 24 out of 36 participants "more strongly agreed" with this survey statement in the case of the accelerometer interface compared to the touch screen gesture interface. In other words, the participants ranked their agreement with this statement for each interface using the likert scale, and in the case of 24 out of 36 participants, the participant gave a response at least 1 point higher on the likert scale when comparing the responses for the accelerometer interface to those of the touch screen gesture interface. If the likert scale is measuring agreement, then a higher score is taken to indicate higher agreement.

Therefore our survey indicated that 24/36 participants (66.67%) more strongly agreed that the accelerometer interface is enjoyable than the touch screen gesture interface. We conduct the following hypothesis test at significance level $\alpha = 0.05$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-1.96 <= z <= 1.96$.
Note: the chosen significance level implies a critical value of 1.96.
Computing the test statistic:

$$z = (0.6667 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 2.00$$

Applying the decision rule: $z = 2.00 > 1.96$, therefore we reject the null hypothesis.

Statement S2 results

On page 38 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population would either somewhat agree or agree that the interfaces were easy to learn.

The survey statement "Learning this interface was easy" was answered for each of the three interfaces by each participant using a 5 point likert-scale (1-5). We observed that 29 of 36 participants (80.55%) either somewhat agreed or agreed with this statement in the case of the touch screen gesture interface. We observed that 35 of 36 participants (97.22%) either somewhat agreed or agreed with this statement in the case of the accelerometer interface. We observed that 33 of 36 participants (91.67%) either somewhat agreed or agreed with this statement in the case of the simulated button interface. We provide the hypothesis test for each claim below.

Our survey indicated that 29/36 participants (80.55%) either somewhat agreed or agreed that learning the interface was easy in the case of the touch screen gesture interface. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.
Note: the chosen significance level implies a critical value of 2.575.

Computing the test statistic:

$$z = (0.8055 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 4.27$$

Applying the decision rule: $z = 4.27 > 2.575$, therefore we reject the null hypothesis.

Our survey indicated that 35/36 participants (97.22%) either somewhat agreed or agreed that learning the interface was easy in the case of the accelerometer interface. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.
Note: the chosen significance level implies a critical value of 2.575.
Computing the test statistic:

$$z = (0.9722 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 5.67$$

Applying the decision rule: $z = 5.67 > 2.575$, therefore we reject the null hypothesis.

Our survey indicated that 33/36 participants (91.67%) either somewhat agreed or agreed that learning the interface was easy in the case of the simulated button interface. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.
Note: the chosen significance level implies a critical value of 2.575.
Computing the test statistic:

$$z = (0.9167 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 5.00$$

Applying the decision rule: $z = 5.00 > 2.575$, therefore we reject the null hypothesis.

Performance

On page 38 we stated that we could say with statistical significance (p-value 0.04) that the majority of the population would persist longer in the game playing with the accelerometer interface than the other interfaces.

We were unable to verify this claim at p-value 0.04. We believe the original calculation may have been completed with the proportion of participants rounded to 67% rather than 66.67%, leading to this error. Whatever the case, we conduct a new hypothesis test (p-value 0.05).

Our survey indicated that 24/36 participants (66.67%) lasted longer in the game playing with the accelerometer interface than the other interfaces. We conduct the following hypothesis test at significance level $\alpha = 0.05$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-1.96 <= z <= 1.96$.

Note: the chosen significance level implies a critical value of 1.96.

Computing the test statistic:

$$z = (0.6667 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 2.00$$

Applying the decision rule: $z = 2.00 > 1.96$, therefore we reject the null hypothesis.

Accelerometer shots fired

On page 39 we stated that we could say with statistical significance (p-value 0.001) that the majority of the population would fire more shots per second with the accelerometer interface than the touch screen gesture interface.

The scroll shooter game kept track of how many shots were fired when a participant used each interface, giving us the ability to calculate how many participants fired more shots per second while using one interface in comparison to another.

Our data indicated that 29/36 participants (80.55%) fired more shots per second with the accelerometer interface than the touch screen gesture

interface. We conduct the following hypothesis test at significance level $\alpha = 0.001$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-3.27 <= z <= 3.27$.
Note: the chosen significance level implies a critical value of 3.27.
Computing the test statistic:

$$z = (0.6667 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 4.27$$

Applying the decision rule: $z = 4.27 > 3.27$, therefore we reject the null hypothesis.

Touch gesture diagonal movement

On page 40 we stated that we could say with statistical significance (p-value 0.001) that the majority of the population will spend more time moving diagonally with the touch gesture interface than other user interfaces.

The scroll shooter game kept track of how much time the participant spent moving their player ship when playing with each interface. As a result, we are able to calculate how many participants spent more time moving in a diagonal direction in one interface compared to another.

Our data indicated that 35/36 participants (97.22%) spent more time moving diagonally with the touch gesture interface than the accelerometer interface. We conduct the following hypothesis test at significance level $\alpha = 0.001$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-3.27 <= z <= 3.27$.
Note: the chosen significance level implies a critical value of 3.27.
Computing the test statistic:

$$z = (0.9722 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 5.67$$

Applying the decision rule: $z = 5.67 > 3.27$, therefore we reject the null hypothesis.

Our data indicated that 29/36 participants (80.55%) spent more time moving diagonally with the touch gesture interface than the simulated button interface. We conduct the following hypothesis test at significance level $\alpha = 0.001$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-3.27 <= z <= 3.27$.
Note: the chosen significance level implies a critical value of 3.27.
Computing the test statistic:

$$z = (0.6667 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 4.27$$

Applying the decision rule: $z = 4.27 > 3.27$, therefore we reject the null hypothesis.

Position touch gesture 4-9

On page 41 we stated that we could say with statistical significance (p-value 0.001) that the majority of the population will spend more time with their player ship in positions 4-9 when using the touch screen gesture interface than with the accelerometer interface, and with statistical significance (p-value 0.01) that the majority of the population will spend more time with their player ship in positions 4-9 when using the touch gesture interface than with the simulated button interface.

The scroll shooter game kept track of how much time the player kept the player ship in each position (1-9) in the game. As a result, we are able to

calculate what proportion of time was spent in positions 4-9 for each participant using each interface, and make comparisons.

Our data indicated that 32/36 participants (88.88%) spent a greater portion of time with their player ship in positions 4-9 when using the touch screen gesture interface than the accelerometer interface. We conduct the following hypothesis test at significance level $\alpha = 0.001$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-3.27 <= z <= 3.27$.

Note: the chosen significance level implies a critical value of 3.27.

Computing the test statistic:

$$z = (0.8888 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 4.67$$

Applying the decision rule: $z = 4.67 > 3.27$, therefore we reject the null hypothesis.

Our data indicated that 26/36 participants (72.22%) spent a greater portion of time with their player ship in positions 4-9 when using the touch screen gesture interface than the simulated button interface. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.

Note: the chosen significance level implies a critical value of 2.575.

Computing the test statistic:

$$z = (0.7222 - 0.5)/\sqrt{0.5 * (1 - 0.5)/36} = 2.67$$

Applying the decision rule: $z = 2.67 > 2.575$, therefore we reject the null hypothesis.

2.7.5 Binomial distribution hypothesis test details

For large values of n it can be argued that the shape of the binomial distribution is close enough to normal that the normal distribution can be used as an approximation.

We have found various guidelines for when a binomial distribution is close enough to normal that a normal distribution can be used as an approximation. For example, $np >= 10$ and $n(p-1) >= 10$ have been suggested as appropriate conditions for using a z-test, or that $np >= 5$ and $n(p-1) >= 5$ are appropriate conditions, or that $n > 30$ is an appropriate condition.

Given the subjectivity as to what value of n is large "enough", and the fact that using a z-test at all entails an approximation, we conduct additional hypothesis testing using the binomial distribution for the sake of providing additional analysis.

Accelerometer interface preference

On page 35 we stated that we could say with statistical significance (p-value 0.02) that the majority of the population most prefers the accelerometer interface, based on the results of a z-test hypothesis test.

The decision rule for this z-test was particularly close ($z = 2.33 > 2.3267$), see section 2.7.4 for the full hypothesis test. We are unable to reject the null hypothesis using a binomial hypothesis test at significance level $\alpha = 0.02$, but are able to do so at significance level $\alpha = 0.03$.

Our survey indicated that 25/36 participants (69.44%) preferred the accelerometer interface. We conduct the following hypothesis test at significance level $\alpha = 0.03$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who preferred the accelerometer interface, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.03$.
Computing the test statistic:

$$pval = 2 * 0.0144 = 0.0288$$

Applying the decision rule: $pval = 0.0288 < 0.03$, therefore we reject the null hypothesis.

Statement S1 results

On page 37 we stated that we could say with statistical significance (p-value 0.05) that the majority of the population would more strongly agree that the accelerometer interface is enjoyable than either of the two other interfaces.

The survey statement "Using this interface was enjoyable" was answered for each of the three interfaces by each participant using a 5 point likert-scale (1-5). We observed that 27 out of 36 participants "more strongly agreed" with this survey statement in the case of the accelerometer interface compared to the simulated button interface. In other words, the participants ranked their agreement with this statement for each interface using the likert scale, and in the case of 27 out of 36 participants, the participant gave a response at least 1 point higher on the likert scale when comparing the responses for the accelerometer interface to those of the simulated button interface. If the likert scale is measuring agreement, then a higher score is taken to indicate higher agreement.

Therefore our survey indicated that 27/36 participants (75.00%) more strongly agreed that the accelerometer interface is enjoyable than the simulated button interface. We conduct the following hypothesis test at significance level $\alpha = 0.05$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who more strongly agreed that the accelerometer interface is enjoyable than the simulated button interface, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.05$.
Computing the test statistic:

$$pval = 2 * 0.001966 = 0.003932$$

Applying the decision rule: $pval = 0.003932 < 0.05$, therefore we reject the null hypothesis.

We observed that 24 out of 36 participants "more strongly agreed" with this survey statement in the case of the accelerometer interface compared to the touch screen gesture interface. In other words, the participants ranked their agreement with this statement for each interface using the likert scale, and in the case of 24 out of 36 participants, the participant gave a response at least 1 point higher on the likert scale when comparing the responses for the accelerometer interface to those of the touch screen gesture interface. If the likert scale is measuring agreement, then a higher score is taken to indicate higher agreement.

The decision rule for this z-test was particularly close ($z = 2.00 > 1.96$), see section 2.7.4 for the full hypothesis test. We are unable to reject the null hypothesis using a binomial hypothesis test at significance level $\alpha = 0.05$, but are able to do so at significance level $\alpha = 0.07$.

Our survey indicated that 24/36 participants (66.67%) more strongly agreed that the accelerometer interface is enjoyable than the touch screen gesture interface. We conduct the following hypothesis test at significance level $\alpha = 0.07$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who more strongly agreed that the accelerometer interface is enjoyable than the touch screen gesture interface, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.07$.
Computing the test statistic:

$$pval = 2 * 0.0326 = 0.0652$$

Applying the decision rule: $pval = 0.0652 < 0.07$, therefore we reject the null hypothesis.

Statement S2 results

On page 38 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population would either somewhat agree or agree that the interfaces were easy to learn.

The survey statement "Learning this interface was easy" was answered for each of the three interfaces by each participant using a 5 point likert-scale (1-5). We observed that 29 of 36 participants (80.55%) either somewhat agreed or agreed with this statement in the case of the touch screen gesture interface. We observed that 35 of 36 participants (97.22%) either somewhat agreed or agreed with this statement in the case of the accelerometer interface. We observed that 33 of 36 participants (91.67%) either somewhat agreed or agreed with this statement in the case of the simulated button interface. We provide the hypothesis test for each claim below.

Our survey indicated that 29/36 participants (80.55%) either somewhat agreed or agreed that learning the interface was easy in the case of the touch screen gesture interface. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who either somewhat agreed or agreed that learning the interface was easy in the case of the touch screen gesture interface, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.01$.

Computing the test statistic:

$$pval = 2 * 0.0001562 = 0.0003124$$

Applying the decision rule: $pval = 0.0003124 < 0.01$, therefore we reject the null hypothesis.

Our survey indicated that 35/36 participants (97.22%) either somewhat agreed or agreed that learning the interface was easy in the case of the accelerometer interface. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who either somewhat agreed or agreed that learning the interface was easy in the case of the accelerometer interface, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.01$.

Computing the test statistic:

$$pval = 2 * 5.38 \times 10^{-10} = 1.08 \times 10^{-9}$$

Applying the decision rule: $pval = 1.08 \times 10^{-9} < 0.01$, therefore we reject the null hypothesis.

Our survey indicated that 33/36 participants (91.67%) either somewhat agreed or agreed that learning the interface was easy in the case of the simulated button interface. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who either somewhat agreed or agreed that learning the interface was easy in the case of the simulated button interface, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.01$.

Computing the test statistic:

$$pval = 2 * 1.136 \times 10^{-7} = 2.272 \times 10^{-7}$$

Applying the decision rule: $pval = 2.272 \times 10^{-7} < 0.01$, therefore we reject the null hypothesis.

Performance

On page 38 we stated that we could say with statistical significance (p-value 0.04) that the majority of the population would persist longer in the game playing with the accelerometer interface than the other interfaces.

The decision rule for this z-test was particularly close ($z = 2.00 > 1.96$), see section 2.7.4 for the full hypothesis test. We are unable to reject the null

hypothesis using a binomial hypothesis test at significance level $\alpha = 0.05$, but are able to do so at significance level $\alpha = 0.07$.

Our survey indicated that 24/36 participants (66.67%) lasted longer in the game playing with the accelerometer interface than the other interfaces. We conduct the following hypothesis test at significance level $\alpha = 0.07$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who lasted longer in the game playing with the accelerometer interface than the other interfaces, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.07$.
Computing the test statistic:

$$pval = 2 * 0.0326 = 0.0652$$

Applying the decision rule: $pval = 0.0652 < 0.07$, therefore we reject the null hypothesis.

Accelerometer shots fired

On page 39 we stated that we could say with statistical significance (p-value 0.001) that the majority of the population would fire more shots per second with the accelerometer interface than the touch screen gesture interface.

The scroll shooter game kept track of how many shots were fired when a participant used each interface, giving us the ability to calculate how many participants fired more shots per second while using one interface in comparison to another.

Our data indicated that 29/36 participants (80.55%) fired more shots per second with the accelerometer interface than the touch screen gesture interface. We conduct the following hypothesis test at significance level $\alpha = 0.001$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who either somewhat agreed or agreed that learning the interface was easy in the case of the touch screen gesture interface, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.001$.

Computing the test statistic:

$$pval = 2 * 0.0001562 = 0.0003124$$

Applying the decision rule: $pval = 0.0003124 < 0.001$, therefore we reject the null hypothesis.

Touch gesture diagonal movement

On page 40 we stated that we could say with statistical significance (p-value 0.001) that the majority of the population will spend more time moving diagonally with the touch gesture interface than other user interfaces.

The scroll shooter game kept track of how much time the participant spent moving their player ship when playing with each interface. As a result, we are able to calculate how many participants spent more time moving in a diagonal direction in one interface compared to another.

Our data indicated that 35/36 participants (97.22%) spent more time moving diagonally with the touch gesture interface than the accelerometer interface. We conduct the following hypothesis test at significance level $\alpha = 0.001$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who spent more time moving diagonally with the touch gesture interface than the accelerometer interface, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.001$.

Computing the test statistic:

$$pval = 2 * 5.38 \times 10^{-10} = 1.08 \times 10^{-9}$$

Applying the decision rule: $pval = 1.08 \times 10^{-9} < 0.001$, therefore we reject the null hypothesis.

Our data indicated that 29/36 participants (80.55%) spent more time moving diagonally with the touch gesture interface than the simulated button interface. We conduct the following hypothesis test at significance level $\alpha = 0.001$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who spent more time moving diagonally with the touch gesture interface than the simulated button interface, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.001$.
Computing the test statistic:

$$pval = 2 * 0.0001562 = 0.0003124$$

Applying the decision rule: $pval = 0.0003124 < 0.001$, therefore we reject the null hypothesis.

Position touch gesture 4-9

On page 41 we stated that we could say with statistical significance (p-value 0.001) that the majority of the population will spend more time with their player ship in positions 4-9 when using the touch screen gesture interface than with the accelerometer interface, and with statistical significance (p-value 0.01) that the majority of the population will spend more time with their player ship in positions 4-9 when using the touch gesture interface than with the simulated button interface.

The scroll shooter game kept track of how much time the player kept the player ship in each position (1-9) in the game. As a result, we are able to calculate what proportion of time was spent in positions 4-9 for each participant using each interface, and make comparisons.

Our data indicated that 32/36 participants (88.88%) spent a greater portion of time with their player ship in positions 4-9 when using the touch screen gesture interface than the accelerometer interface. We conduct the following hypothesis test at significance level $\alpha = 0.001$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who spent a greater portion of time with their player ship in positions 4-9 when using the touch screen gesture interface than the accelerometer interface, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.001$.
Computing the test statistic:

$$pval = 2 * 9.70 \times 10^{-7} = 1.94 \times 10^{-6}$$

Applying the decision rule: $pval = 1.94 \times 10^{-6} < 0.001$, therefore we reject the null hypothesis.

Our data indicated that 26/36 participants (72.22%) spent a greater portion of time with their player ship in positions 4-9 when using the touch screen gesture interface than the simulated button interface.

The decision rule for this z-test was particularly close ($z = 2.67 > 2.575$), see section 2.7.4 for the full hypothesis test. We are unable to reject the null hypothesis using a binomial hypothesis test at significance level $\alpha = 0.01$, but are able to do so at significance level $\alpha = 0.02$.

We conduct the following hypothesis test at significance level $\alpha = 0.02$.
Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who spent a greater portion of time with their player ship in positions 4-9 when using the touch screen gesture interface than the simulated button interface, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.02$.
Computing the test statistic:

$$pval = 2 * 0.00566 = 0.01132$$

Applying the decision rule: $pval = 0.01132 < 0.02$, therefore we reject the null hypothesis.

Chapter 3

Computer Science App Study

The following work was published in the proceedings for the Gamification 2013 conference:

K. Browne and C. Anand, Gamification and serious game approaches for introductory computer science tablet software, Proceedings of the First International Conference on Gameful Design, Research, and Applications. ACM, 2013.

The work has been printed in this thesis under license (ACM Publishing License and Audio/Video Release).

3.1 Abstract

In this paper, we overview the design of tablet apps built to teach introductory computer science concepts, and present the results and conclusions from a study conducted during a first year computer science course at McMaster University. Game design elements were incorporated into the apps we designed to teach introductory computer science concepts, with the primary aim of increasing student satisfaction and engagement. We tested these apps with students enrolled in the course during their regular lab sessions and collected data on both the usability of the apps and the student's understanding of the concepts. Though overall we found students preferred instruction with the apps compared to more traditional academic instruction, we found that students also recommended combined instruction using both traditional methods and the apps in the future. Based on this we conclude that gamification and serious game design approaches are effective at increasing student satisfaction, and make several recommendations regarding the usage and design of educational software incorporating game design elements.

3.2 Introduction

The first year course CS1MA3 was the introductory computer science course in the Department of Computing & Software at McMaster University. The course was a breadth approach to introductory computer science, with students taught different algorithms and how to derive their own algorithms, the basics of functional programming, binary numbers, graphs, and the basics of CPUs and assembly language. All of the ideas are presented with the overall goal of developing the students' problem solving skills in the context of computer science.

Though the course was recommended for students wishing to pursue computer science in upper years, students from a diverse selection of backgrounds took the course as an elective. In past years, many students have found themselves unprepared for what a "computer course" really entails at the university level. Many students have never been exposed to basic computer science, have never written a computer program or have been taught only the mechanical aspects of programming. Such students experience frustration and dissatisfaction when they are expected to solve new types of problems and use new ways of thinking. This dissatisfaction often leads to disengagement during instruction, and failure to advance in the program.

The authors of this work are also involved in an outreach program at McMaster University that visits elementary school classrooms with activities meant to give students a fun and engaging example of computer science. One of the most successful activities in this program is one that we developed internally; what made this activity successful was how well it engaged students. The activity involved having the students write computer code to fix a simple table tennis video game that was initially presented to the students as being 'broken' (i.e. balls would fly through paddles and off the screen because the correct code was not yet present). Our observation and intuition suggested it was the table tennis game element that led to this higher satisfaction and engagement. We were motivated by these observations to try and reproduce this higher satisfaction by introducing educational software containing game design elements into the instruction of CS1MA3.

Our primary aim in conducting this study was to investigate whether integrating educational software incorporating games and/or game design elements into the instruction of introductory computer science could increase student satisfaction and engagement relative to traditional instructional methods. Secondary aims included examining what specific game design features of the apps are effective at increasing satisfaction, whether using the apps leads to a better understanding of the related concepts, and how such educational

software can best be integrated into instruction of the concepts.

As a result we designed and created the apps featured in this study, empirically tested them during the Fall 2012 iteration of CS1MA3, and present the results here in a manner in which we hope is reproducible for others teaching these and similar concepts. Based on the strong endorsement of the study participants, we will continue to use all of the relevant applications in the successor course, CS1JC3 (Computational Thinking). We have also used the study feedback to improve the game elements of the binary numbers app, which is currently being used successfully with over 2000 elementary school students as part of the NSERC PromoSCIENCE-supported outreach workshop "Software: Tool for Change".

In this section we will first overview some terminology and related work.

The potential for tablets as a learning tool is an active area of research and the positive observations reported thus far [CG11] in using tablets in the classroom, and more importantly the potential for contributing further findings to the existing literature motivated our decision to use iPads for this study.

The "usage of game design elements to motivate user behaviour in non-game contexts" is known as gamification, and is considered by Deterding et al to be a concept warranting deeper empirical exploration [Det11]. Deterding et al define a serious game to be "a full fledged game for non-entertainment purposes", where as "gamified applications merely incorporate elements of games" and acknowledge that "the boundary between a 'game' and an 'artifact with game elements' can often by blurry". Our apps incorporate game design elements to varying degrees, and as such our work by its nature involves some examination of the intersection between gamification and serious games.

There is little literature at present covering gamification and / or serious game approaches to educational tablet software, but what work is available shows encouraging results [WCD11, Yan11]. Popular educational websites focusing on interactive online programming that use gamification in their design include Codecademy and Khan Academy. However, both websites are focused on teaching computer programming, where as our apps are focused on teaching other computer science concepts.

3.3 Application Design

In this section we will overview the six apps designed to teach binary search, binary numbers, CPU / assembly language, polynomial graphs, quicksort and Dijkstra's algorithm. A common design element across all of the apps is the incorporation of interactive representations of the relevant concepts. Game

elements such as objectives, rewards, penalties, levels, narrative, multiplayer, and in-game assistance are incorporated into the apps to varying degrees. We define in-game assistance to occur when the app provides corrective help or feedback after the user has failed to make the correct actions.

3.3.1 Binary Search App

The binary search app shown in Figure 3.1 begins with a row of coconuts on the bottom half of the screen, and a message telling the user to 'crack the coconuts to find the golden egg'. When the user taps on a coconut an arrow pointing either left or right appears, depending on which direction the coconut is located. If the user does not follow the correct sequence of taps for binary search, it will be suggested to the user that there is another way of finding the golden egg.

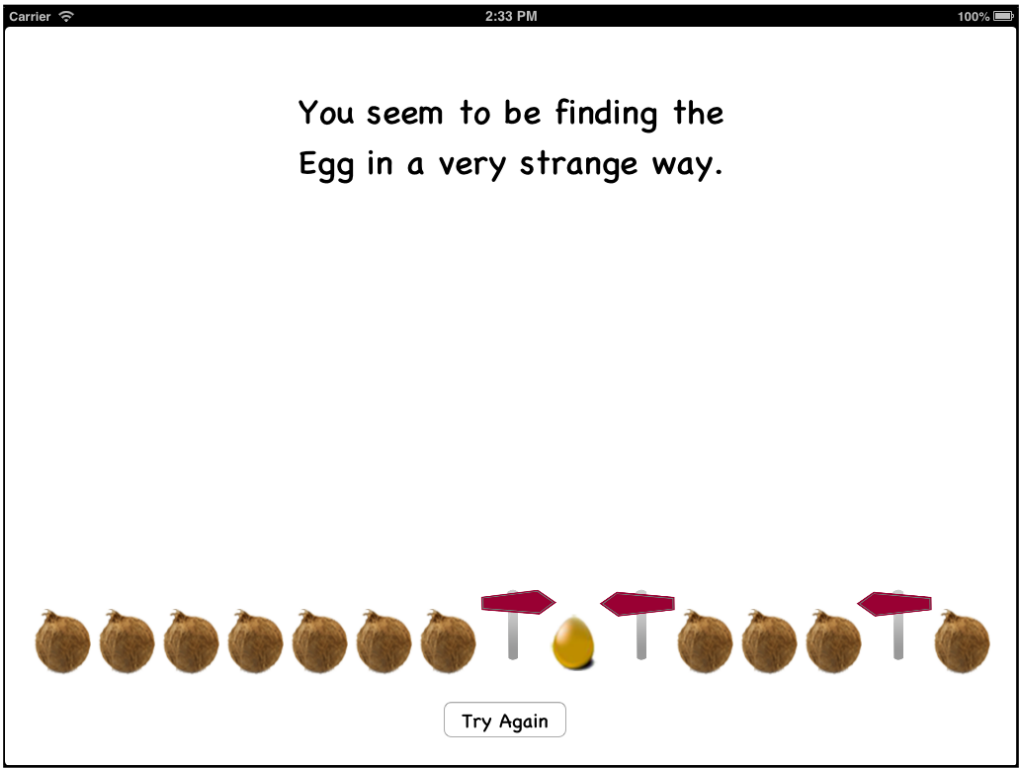

Figure 3.1: Binary search app

If the user follows the correct sequence of taps for binary search 5 times in a row, they will be rewarded with a treasure chest that they can open by

tapping. When the user taps on the treasure chest they are treated to a screen that tries to explain why a maximum of 4 taps are required to reach the golden egg by showing the user the binary search tree. The app goes on to pan across a series of alternative search trees before exiting automatically.

3.3.2 Binary Number App

The binary number app shown in Figure 3.2 uses both animations of binary number conversion and arithmetic as well as a multiplayer network game to assist in the instruction of binary number concepts. The app contains three different pages: conversion, arithmetic, and more (which contains the game).

The conversion page allows the user to select a decimal number using a slider on the bottom of the screen, and the number is shown converted live into binary at the top of the screen. By clicking one of two arrows the user is able to see an animation of the number being converted from one base to another in either direction.

The arithmetic page allows the user to modify two binary numbers by clicking each digit back and forth between '0' and '1'. The user can then select either addition or subtraction and play an animation of the arithmetic taking place.

If the iPad is connected to a wifi network with other iPads using the app, the game screen allows the player to search for others to play the game. The game involves both players first tapping spaces in a grid to represent different binary numbers. The game then sends the decimal numbers that were created in the process to the other player for them to convert into the correct grid. The first player to convert the other player's numbers wins the game.

3.3.3 CPU App

The CPU app involves a single screen containing a visual representation of memory, registers, operators and a set of instructions as shown in Figure 3.3. The users execute each instruction by performing actions such as dragging data from memory to registers or registers to memory. In the case of operations, users drag the appropriate operation to an arithmetic logic unit together with register arguments before the result is computed, and must be dragged into the appropriate register.

Correct actions advance the user to the next instruction along with playing an approving sound effect as a reward, and incorrect actions result in

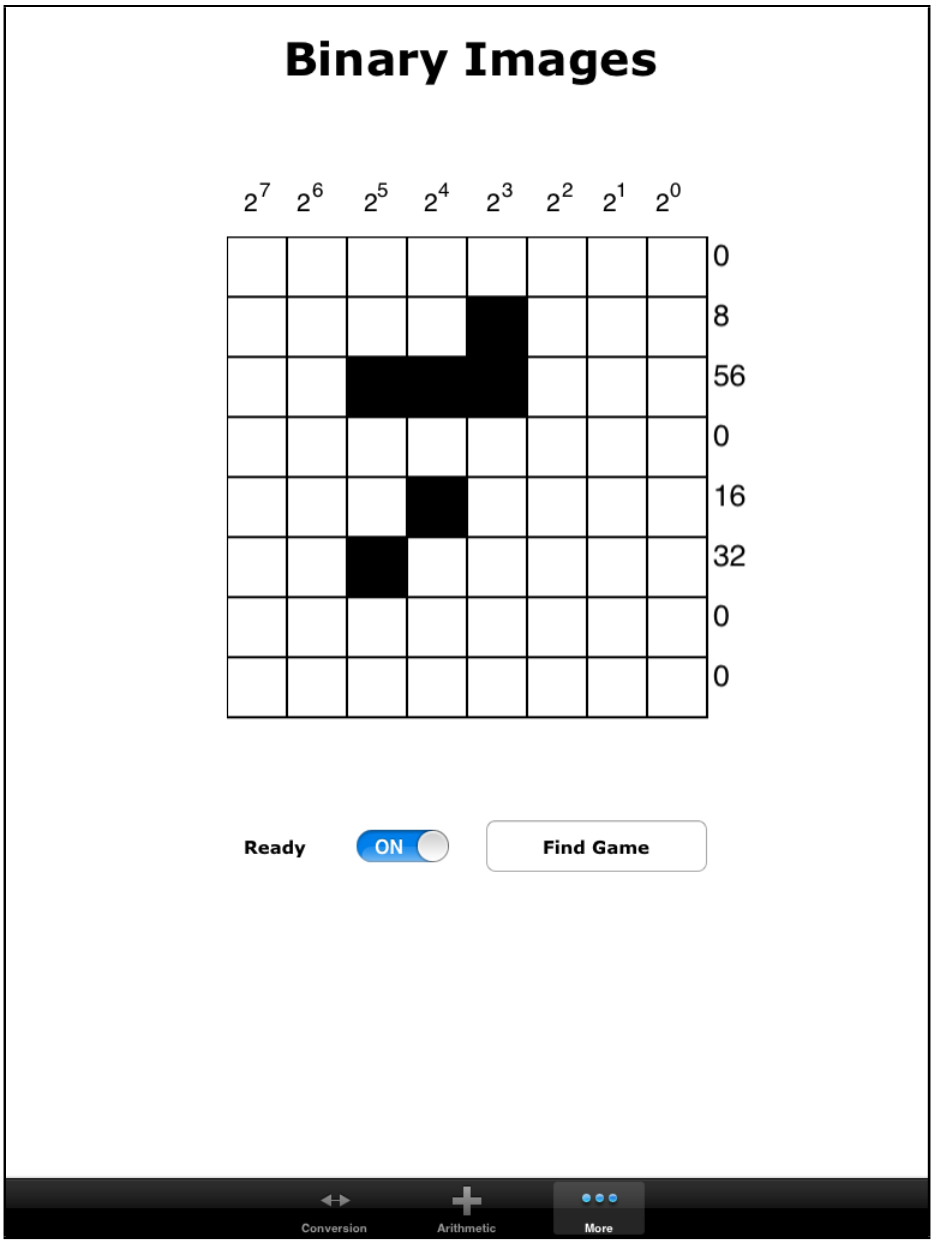

Figure 3.2: Binary number app

a disapproving sound effect as a penalty. There are 3 'levels' total consisting of 3 different sets of instructions for the user to work through.

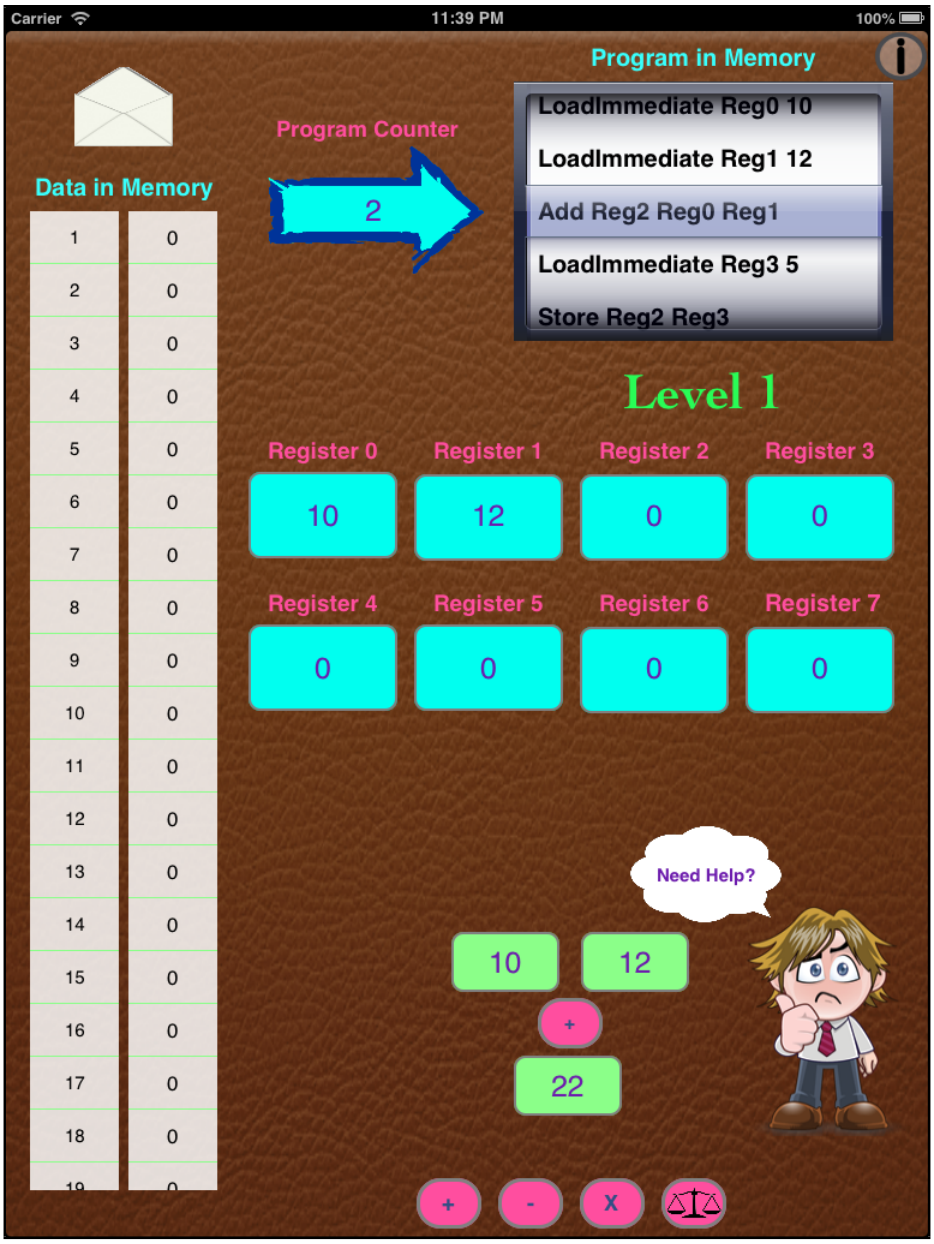

Figure 3.3: CPU app

If at any point the user hasn't made any sort of action within a few seconds, a 'Need Help?' cartoon person appears. By clicking on this the

app will animate the correct action automatically and advance to the next instruction.

3.3.4 Polynomial App

The polynomial app shown in Figure 3.4 begins with a blank screen and a one line textbox that users are told to tap. When the user taps on the textbox they are presented a calculator-style menu which allows the user to enter a polynomial.

After the user enters the polynomial they can create the associated graph by dragging the operators, numbers and symbols down into the blank space. When an operator is dragged into the space below there are blank nodes connected to either side to allow the user to build a tree by connecting more operators, numbers or symbols as leaves. If the user makes a mistake they can drag the tree into the garbage bin to start over. If the user wants an entirely new polynomial expression they can tap the 'reset' button. The user can optionally turn on additional parenthesis to outline the order of operations.

The user can check whether the polynomial graph is correct or not by tapping a 'check' button. If the graph is incorrect the user will receive a red X, and if the graph is correct the user will receive a green check mark instead.

3.3.5 Quicksort App

The quicksort app shown in Figure 3.5 allows the user to step through a demonstration and explanation of quicksort by selecting pivot points and tapping Next. The app starts off with a list of 5 unordered numbers and the user is told to select a pivot upon tapping the 'Start/Next' button. Depending on the pivot that the user selects, by clicking on Next repeatedly the app explains step-by-step the swaps that take place in between the need to select a new pivot.

As the list is broken into sublists to which quicksort can be applied, the graph of arrows in the space below is filled in as required. A red X is used to represent empty sublists. The user can click 'Reset' at anytime to start off with a new unordered list.

3.3.6 Dijkstra App

The Dijkstra app shown in Figure 3.6 starts off with a mostly blank screen that the user is told to tap in order to place the nodes. The first time the user taps the screen a node in the shape of the McMaster University engineering

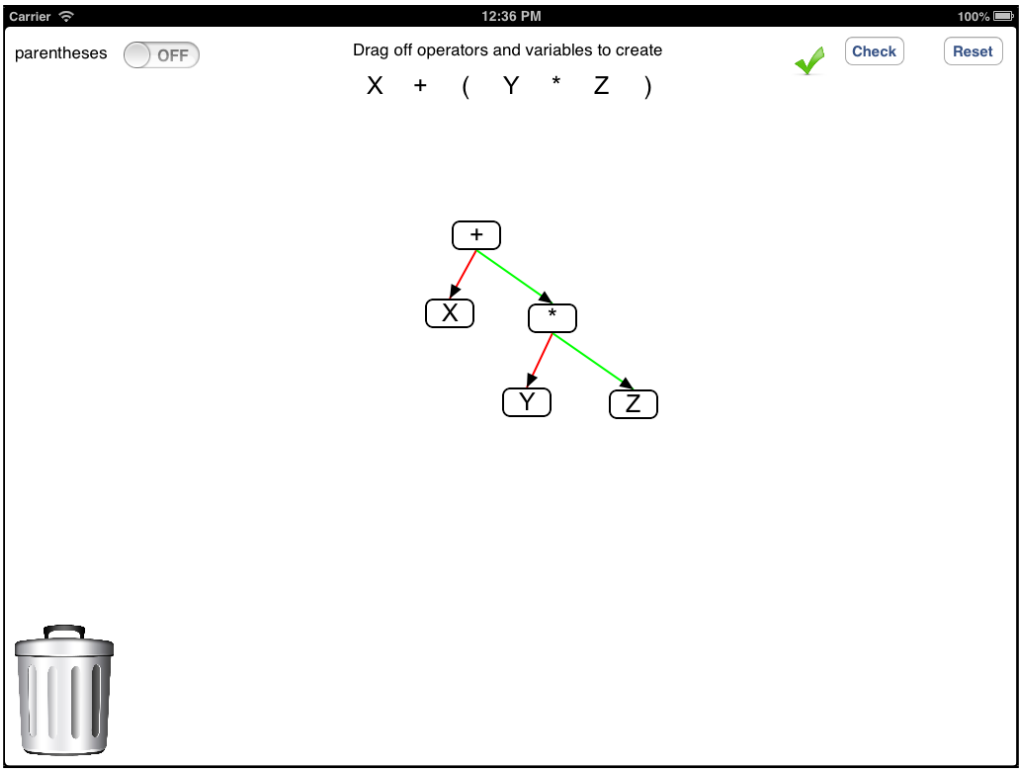

Figure 3.4: Polynomial app

department 'fireball' logo is placed as the initial node. The user can continue to tap the screen at different positions to place more nodes. The only restriction on the number of nodes that the user can place is how closely nodes can be placed; the app will just not place down a node if the user taps too closely to an existing node.

Once the user is finished placing the nodes, they can use 'Start', 'Back' and 'Next' buttons to step through a graphical trace of the algorithm's execution on the graph that they just created. A node that has been visited is marked green, while an eye cartoon and a blue path in the direction of a node are used to indicate the current node and unvisited nodes which are having their tentative distances calculated.

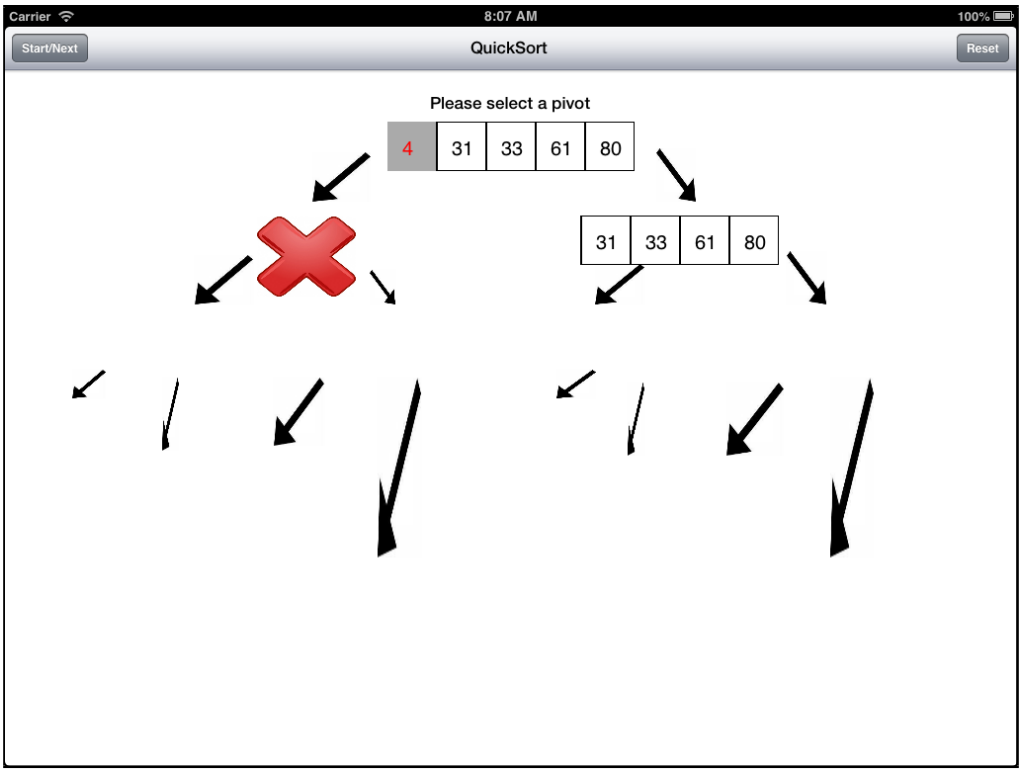

Figure 3.5: Quicksort app

3.3.7 App Comparison

The different game design elements incorporated into the apps are presented in Table 3.1. Notably the quicksort and Dijkstra apps do not contain any game design elements, other than interactive representations of the concepts, a design approach common to all of the apps. We consider an objective to be a goal presented by the app that the user is expected to complete. The polynomial, quicksort and Dijkstra apps do not contain explicit objectives, but rather allow for free form improvisational experiences with the concepts. The binary search, binary numbers and CPU apps all do contain explicit objectives or goals that create a more proper *game*.

In the case of the binary number app, only a subsection of the app contains a game with explicit objectives, with the other two subsections allowing for improvisational experiences with the concepts. As such, the binary search

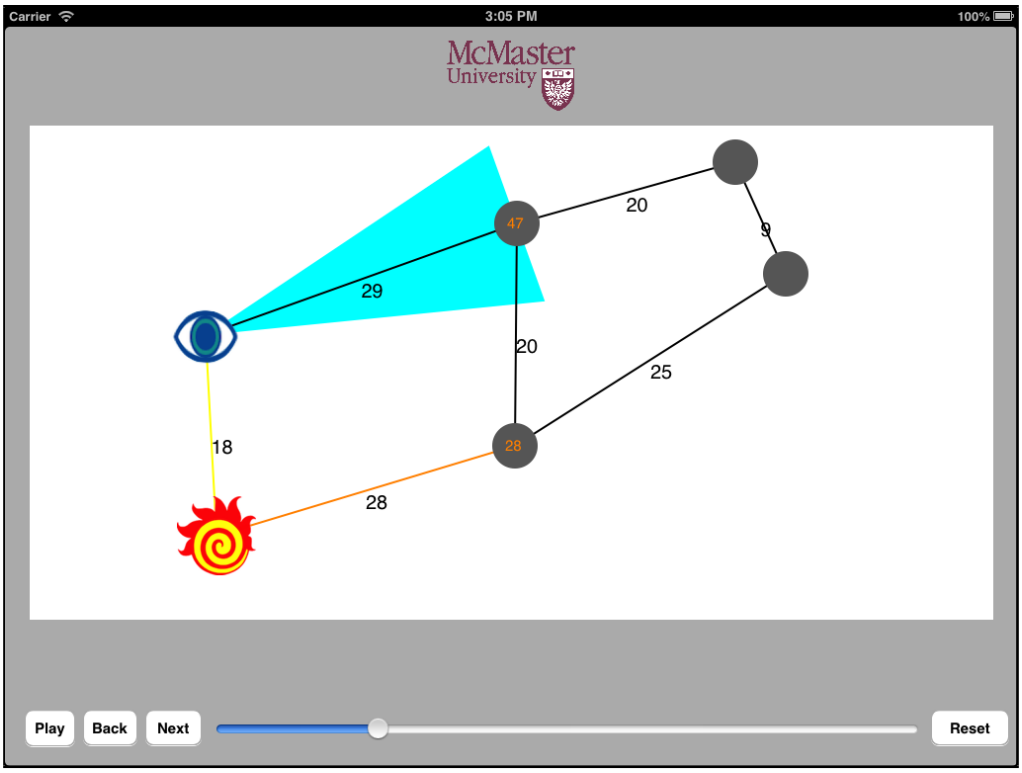

Figure 3.6: Dijkstra app

and CPU apps may be considered serious games, with the binary numbers app containing a serious game, and the remainder of the apps more accurately described as instances of gamification rather than serious games.

3.4 Experiment Design

In addition to attending lecture sessions, CS1MA3 students each attend a weekly lab 50 minutes in duration. Six weeks of lab sessions were set aside during the course for experiment sessions using the six iPad apps, with each of these six weeks set aside for an experiment with a specific iPad app. Each of the six weeks of sessions itself contained six sessions total; participants were assigned to go to a specific session each week. All sessions took place in the

App	Game Design Elements			
	Objectives	**Rewards**	**Penalties**	**Levels**
Binary Search	X	X		
Binary Numbers	X			
CPU	X	X	X	X
Polynomial		X	X	
Quicksort				
Dijkstra				

App	Game Design Elements		
	Narrative	**Multiplayer**	**Assistance**
Binary Search	X		X
Binary Numbers		X	
CPU			X
Polynomial			
Quicksort			
Dijkstra			

Table 3.1: Game design elements incorporated into each iPad app.

same computer lab at McMaster University. Nine iPads were available during these sessions, requiring some students to share devices.

Students were given the option of participating in the study or not. A code system and consent process managed by a third party was used so that the instructors and teaching assistants of CS1MA3 would never know which students participated in the study and which did not.

3.4.1 Pre-Experiment Survey

The following information was gathered from participants with a pre-experiment questionnaire at the beginning of the term: gender (Male/Female/Other), handedness (Right/Left/Ambidextrous), years of study completed at the college or university level (1- 20), year of study in current undergraduate program

(1-5), current undergraduate program.

The participants were also asked to rate their overall performance in any past mathematics courses at any level of study from 1-10, with 1 being "failing performance overall" and 10 being "perfect performance overall". They were asked to do the same for their overall performance in any past computer science courses at any level.

The participants were also asked to rate their expertise with the following different media and interfaces: mobile phones, mobile video games, iPad / other tablets - general usage, iPad / other tablets - video games, and touch screen interfaces. Participants rated their expertise by selecting one of the following expertise levels, based on descriptions of the expertise levels: 'no expertise', 'some expertise', 'typical expertise', 'above average expertise', and 'expert'. For analysis purposes these expertise levels were assigned the numeric values 1-5 from no expertise to expert.

3.4.2 Experiment Session Protocol

The following procedure was followed during each experiment session:

1. Students were given a short (i.e. 2 minutes) quiz to fill out at the start of class.

2. Students were then given either 20 minutes of instruction in the relevant concept with usage of the apps, or 20 minutes of instruction without usage of the apps (i.e. "traditional instruction"). Half of the tutorials started with traditional instruction, and half started with the apps being used. Over the six experiment weeks we alternated which labs started with which instructional method.

3. Students were given the same quiz. If students had just finished using the apps to learn the concepts, they were also given a usability questionnaire.

4. Students in labs which started by using the apps were given 20 minutes of instruction without using the apps, and vice versa.

5. Students were given the same quiz. If students had just finished using the apps to learn the concepts, they were also given a usability questionnaire. Students were given a post-experiment questionnaire.

3.4.3 Usability Survey

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

- **S1** The software was easy to use.
- **S2** It was difficult to learn how to use the software.
- **S3** I enjoyed using this software.
- **S4** I thought there was too much inconsistency in this software.
- **S5** The touchscreen finger gestures required to manipulate the software felt natural.
- **S6** The software was too slow to use efficiently.
- **S7** The graphics in the software were effective.
- **S8** I found the iPad uncomfortable to hold while using the software.
- **S9** The software helped me to learn the concepts.
- **S10** I felt confused trying to understand the concepts with this software.
- **S11** I feel confident that I understand the concepts.
- **S12** I would not recommend that others try to learn the concepts by using this software.

The participants could choose from: strongly disagree, somewhat disagree, neutral, somewhat agree, and strongly agree. The design of the usability survey was inspired by features of the System Usability Scale (SUS) [BKM08]. In a similar manner to SUS a usability score can be calculated, by assigning the average response of each statement a score contribution from 0-4. The odd numbered statement's score contribution is a value of 0-4 from strongly disagree to strongly agree, and the even numbered statement's score contribution is a value of 4-0 from strongly disagree to strongly agree. Dividing the sum of the scores by 48 and multiplying the result by 100 gives a score with a range from 0 to 100.

The usability survey also contained two questions, each with a blank space for feedback: "What did you like about the software?" and "What didn't you like about the software?".

3.4.4 Post-Experiment Questionnaire

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

1. Did you prefer learning the concept(s) taught during the tutorial with the educational software or without the educational software? (circle one)

2. Why was this your preference? *(blank space)*

3. Based on your experience, next year when we teach the concepts taught in this tutorial should we only use the educational software, only use traditional tutorial methods (such as lecturing), or use both? (circle one)

3.4.5 Quizzes

A quiz was developed for each of the six weeks of experiment sessions, with the same quiz being administered three times during each experiment session with the goal of measuring learning as it took place over the session. The quiz questions were short answer form and meant to be easy to quickly complete if the participant understood the relevant concept. Quizzes were marked in a pass-fail manner such that only quizzes completed 100% correctly were considered a pass. The following types of questions were asked on each quiz:

- The binary search quiz required students to describe the algorithm.
- The binary number quiz required students to perform binary number conversions.
- The CPU quiz required students to trace the execution of assembly instructions given a table for register values.
- The polynomial quiz required students to draw an expression graph for a given polynomial.
- The quicksort quiz required students to describe the algorithm.
- The Dijjkstra quiz required students to determine what order Dijkstra's algorithm visited nodes in a given weighted graph.

3.4.6 Traditional Instruction

Traditional instruction involved Kevin Browne giving the students a presentation which featured an explanation of the concept(s) followed by examples to illustrate them. In all cases the presentation was followed by a Q&A session with the students to help ensure they understood the concepts.

3.5 Results and Discussion

In total 101 CS1MA3 students registered to participate in the study and 6 students did not not. The number of participants that actually attended to participate in each week of the experiment sessions varied. The participants were made up of 71 males, 29 females and 1 participant who did not report their gender; 88 were right-handed, 8 were left-handed, 4 were ambidextrous and 1 participant did not report their handedness. The compiled results showed no real significant differences between participants of different genders or handedness, so we omit discussion of these differences from this analysis but still provide this data and other pre-experiment questionnaire data for context.

The average years of study completed at the college or university level was 0.73 ($\sigma = 1.19$) and the average year of study in the participant's current program was 1.53 ($\sigma = 1.01$). There were 33 participants registered in computer science, 22 in mathematics, 8 in physical sciences, 8 in life sciences, 4 in social sciences, with the remainder in other subjects including humanities, economics, and multimedia. There were 32 students that reported having taken a college or university level mathematics course before, and 65 participants reported they did not. The average participant rating of their performance overall in any past mathematics course was 7.70 ($\sigma = 1.49$), and in any computer science courses was 7.19 ($\sigma = 2.03$).

The data regarding participant expertise with various media and interfaces which was collected in the pre-experiment questionnaire is presented in Table 3.2.

Interface/Media	**Avg.**	**SD**
Mobile phones	3.93	1.05
Mobile video games	3.16	1.12
iPad / other tablets - general usage	2.95	1.20
iPad / other tablets - video games	2.84	1.19
Touch screen interfaces	3.65	1.08

Table 3.2: Participant expertise data - average and standard deviation

3.5.1 General Results

The binary search app was the first experiment conducted, and many participants vocalized frustration with our post-experiment questionnaire that initially only contained a question asking them whether they preferred traditional instruction or instruction with the apps. Many participants were adamant that using *both* forms of instruction was preferable.

As a result we added a third question to the post-experiment questionnaire asking participants whether in the future these concepts should be taught with only the traditional methods, only the apps, or both. One of the strongest results across the experiments performed with each app was that a strong majority of participants recommended using both methods to teach the relevant concepts in the future, as seen in Table 3.3.

Experiment	**Preferred**		**Recommended**		
	App	**Lesson**	**App**	**Lesson**	**Both**
Binary Search	81% (56)	19% (13)	N/A	N/A	N/A
Binary Numbers	68% (42)	32% (20)	3% (2)	11% (7)	85% (52)
CPU	63% (38)	37% (22)	7% (4)	13% (8)	80% (48)
Polynomial	69% (41)	31% (18)	5% (3)	12% (7)	83% (48)
Quicksort	51% (25)	49% (24)	2% (1)	30% (14)	68% (32)
Dijkstra	69% (34)	31% (15)	4% (2)	14% (7)	82% (40)

Table 3.3: Participant preferences and recommendations

Using z-score confidence intervals for these sample proportions, we can say with 99% confidence that the majority of the population prefers instruction with the binary search, binary numbers, polynomial and Dijkstra apps to traditional instruction, and with 95% confidence in the case of the CPU app. Again using z-score confidence intervals we can say with 99% confidence across all the apps that the majority of the population recommends a combination of the app and traditional instruction be used in the future (except for the binary search app where we did not collect this data).

The usability survey scores obtained using the formula described previously are shown in in Figure 3.7. Interestingly, the usability survey scores follow a very similar pattern to the percentage of participants that preferred instruction with each app; the correlation coefficient between the two is 0.86.

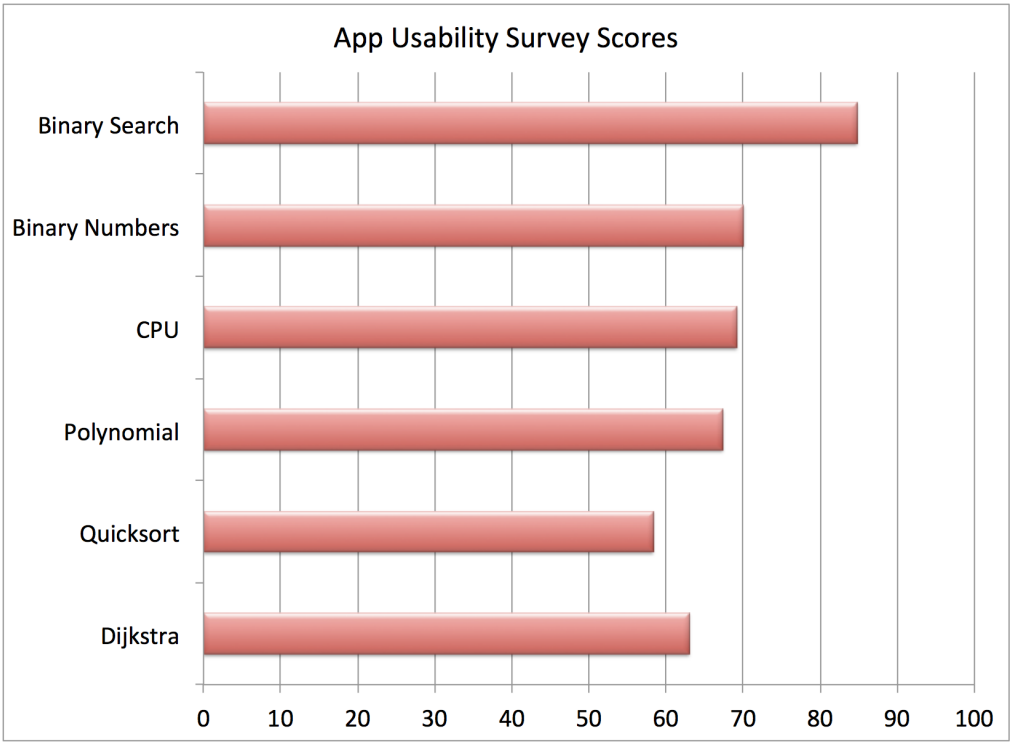

Figure 3.7: App usability survey scores (max = 100)

The quiz results presented in Table 3.4 show the results for the sessions that started with a lesson, the sessions that started with the apps, and the combined results across all sessions. The results are consistent with the majority of participants recommending future instruction using both apps and traditional instruction, as by the completion of quiz 3 participants had experienced both instructional methods and quiz 3 averages were generally higher than those of quiz 2 (which was conducted after participants had experienced only one instructional method).

Participants gave ample feedback both orally and through the post-experiment questionnaire as to what they perceived as the relative merits

Experiment	Order	Quiz 1	Quiz 2	Quiz 3
Binary Search	App-first sessions	25.00	55.56	55.56
	Lesson-first sessions	44.12	73.53	82.35
	All sessions	34.29	64.29	68.57
Binary Numbers	App-first sessions	56.25	75.00	75.00
	Lesson-first sessions	36.67	66.67	70.00
	All sessions	46.77	70.97	72.58
CPU	App-first sessions	3.45	27.59	62.07
	Lesson-first sessions	0.00	54.84	77.42
	All sessions	1.67	41.67	70.00
Polynomial	App-first sessions	3.57	39.29	71.43
	Lesson-first sessions	0.00	31.43	51.43
	All sessions	1.59	34.92	60.32
Quicksort	App-first sessions	0.00	19.35	53.33
	Lesson-first sessions	5.00	70.00	70.00
	All sessions	1.96	39.22	60.00
Dijkstra	App-first sessions	8.00	56.00	80.00
	Lesson-first sessions	8.33	70.83	79.17
	All sessions	8.16	63.27	79.59

Table 3.4: Quiz results (average)

of each type of instruction, as summarized in Table 3.5. There was practically unanimous consensus on the idea that while the traditional lesson-based method was better for 'explaining' or 'teaching' the concepts, the apps were better for 'practicing' or 're-enforcing' concepts. Most participants who expressed this also suggested that the apps be used *after* the lecture.

3.5.2 Binary Search Experiment

There were 70 participants in the binary search sessions, with 35 in experiment sessions that started off with traditional instruction and 35 that started off with instruction with the app.

Many participants expressed frustration at the end part of the app

Lesson Strengths	**App Strengths**
More comprehensive	More fun
Learner can ask questions	Learner can try the app as many times as needed
Better for explaining concepts	Better for practicing concepts
Better for those who learn through listening	Better for those who learn through doing

Table 3.5: Participant perceptions of relative instructional method strength

where a binary search tree is displayed, as they were unfamiliar with binary search trees or the concept of trees in general at the time they used this app.

The most common positive feedback about the app was that it was 'easy to use', 'fun' and 'interactive'. That the app guided participants to understanding binary search by correcting incorrect sequences of taps was praised by participants who appreciated discovering the concept through trial and error.

3.5.3 Binary Numbers Experiment

There were 62 participants total in the binary number sessions, with 29 in sessions that started off with traditional instruction and 33 that started off with instruction with the app.

A few users expressed frustration regarding a lack of instructions and 'not knowing what to do'. There was some negative feedback regarding the time required to make the multiplayer game work; this was most likely due to difficulty in connecting to the network and finding a partner to play with.

In contrast to the complaints about 'not knowing what to do', 15 participants left positive comment feedback as to how 'easy to use' the app was to use. Similarly the multiplayer game subsection received more positive feedback than negative and was described as being fun and engaging.

3.5.4 CPU Experiment

There were 60 participants total in the CPU sessions, with 31 in sessions that started off with traditional instruction and 29 that started off with instruction with the app.

There was some negative feedback about the app being confusing or hard to use, though some of this was qualified by notes that it was only 'at first'. There was also some annoyance over the beeping sounds that the app would make as participants performed instructions.

The help feature of the app received much praise from participants for preventing them from becoming stuck at any point.

3.5.5 Polynomial Experiment

There were 59 participants total in the polynomial sessions, with 26 in sessions that started off with traditional instruction and 33 that started off with instruction with the app.

The most common negative feedback was the lack of instructions or indication as to how the app worked.

Again any negative feedback as to the difficulty of using the app was largely outnumbered by the 14 positive comments stating that the app was 'easy to use'. Interestingly, despite not containing many explicit game elements, several users still described it as fun or game-like.

3.5.6 Quicksort Experiment

There were 49 participants total in the quicksort sessions, with 20 in sessions that started off with traditional instruction and 29 that started off with instruction with the app.

The quicksort app was the most poorly received app relative to the others due to several issues. One particular point of frustration was the usage of the red X to represent empty sublists, as the participants found this discouraging.

There was also feedback that the app didn't do enough in terms of animations to explain the concepts, or involve the participant enough in terms of giving them something to do. Other negative feedback suggested the app was 'boring' and cited the lack of a game element.

Positive feedback about the app included praise for its simplicity and that it was easy to use.

3.5.7 Dijkstra Experiment

There were 49 participants total in the Dijkstra sessions, with 25 in sessions that started off with traditional instruction and 24 that started off with instruction with the app.

Similar to the polynomial app the most common frustration about the Dijkstra app was the lack of instructions or explanations about how to use the app and what the app was doing.

However, the graphical explanation of the algorithm was widely praised. Others praised the app for its simplicity, the ability to step though the algorithm, and the ability to setup the nodes and paths.

3.6 Conclusion

We conclude that integrating educational software incorporating game design elements into the instruction of introductory computer science does increase student satisfaction and engagement, on the basis of the participants' preference for instruction with the apps and their recommendation that the apps be used as part of future instruction.

Based on the oral and written feedback given to us by the participants and our observations during experiment sessions we can make several recommendations about the effectiveness of specific game design elements. While we recommend that apps contain rewards (e.g. treasure chest, check mark) we do not recommend that apps contain penalties (e.g. disapproving sounds, red X) or even symbols and sounds not meant as penalties but that could be perceived as discouraging. The rewards resulted in positive feedback whereas penalties resulted in frustration. This leads to our second point, rather than penalizing users for incorrect actions as is perhaps more typical in video games made exclusively for entertainment purposes, we recommend that educational software provide in-game assistance in the form of corrective help or feedback. The in-game feedback of the binary search and CPU apps received exclusively positive feedback from participants.

Based on our observations and feedback received from the participants, we suggest that software should augment and enhance, but not replace, traditional instruction. We suggest that software should be used primarily for practicing understanding of concepts, and that it should incorporate game design elements to provide increased user satisfaction.

We believe the results are inconclusive as to whether using the apps resulted in a better understanding of the concepts. For ethical reasons, we did not design our experiments with a control group receiving only traditional instruction—an approach which would have allowed us to compare final exam results. However, based on the overwhelming recommendation to incorporate the apps with traditional instruction in the future, we feel that this was the right decision.

3.7 Bibliography

[BKM08] Aaron Bangor, Philip T. Kortum, and James T. Miller. An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6):574–594, 2008.

[CG11] Alma L Culén and Andrea Gasparini. ipad: a new classroom technology? a report from two pilot studies. *INFuture Proceedings*, pages 199–208, 2011.

[Det11] S. Deterding. Situated motivational affordances of game elements: A conceptual model., 2011. Presented at Gamification: Using Game Design Elements in Non-Gaming Contexts, a workshop at CHI 2011. Retrieved from http://gamification-research.org/wp-content/uploads/2011/04/09-Deterding.pdf.

[WCD11] K. Wattanatchariya, S. Chuchuaikam, and N. Dejdumrong. An educational game for learning wind and gravity theory on ios: Drop donuts. In *Computer Science and Software Engineering (JCSSE), 2011 Eighth International Joint Conference on*, pages 387 –392, may 2011.

[Yan11] Feng Yan. A sunny day: Ann and rons world an ipad application for children with autism. In Minhua Ma, Manuel Fradinho Oliveira, and Joo Madeiras Pereira, editors, *Serious Games Development and Applications*, volume 6944 of *Lecture Notes in Computer Science*, pages 129–138. Springer Berlin / Heidelberg, 2011.

3.7 Amendments

In this section we provide clarification, corrections and additional analysis based on feedback provided by a reviewer.

In Section 3.7.1 we correct our usage of the population parameter symbol. In Section 3.7.2 we provide clarifications about the correlation calculation performed in the paper.

In Section 3.7.3 we clarify our claims about preferences and recommendations of the "majority of the population". In Section 2.7.4 we present the z-test hypothesis tests done in this paper in detail, and in Section 2.7.5 we present binomial distribution hypothesis tests.

In Section 3.7.6 we conduct ANOVA hypothesis tests of the quiz scores as additional analysis.

3.7.1 Population parameter symbol

On page 86 the population parameter symbol σ is used where the sample statisic symbol s should have been used.

1.
   - **Original:** The average years of study completed at the college or university level was 0.73 ($\sigma = 1.19$) and the average year of study in the participant's current program was 1.53 ($\sigma = 1.01$).
   - **Revison:** The average years of study completed at the college or university level was 0.73 ($s = 1.19$) and the average year of study in the participant's current program was 1.53 ($s = 1.01$).

2.
   - **Original:** The average participant rating of their performance overall in any past mathematics course was 7.70 ($\sigma = 1.49$), and in any computer science courses was 7.19 ($\sigma = 2.03$).
   - **Original:** The average participant rating of their performance overall in any past mathematics course was 7.70 ($s = 1.49$), and in any computer science courses was 7.19 ($s = 2.03$).

3.7.2 Correlation calculation

On page 88 we stated that the usability survey scores follow a very similar pattern to the percentage of participants that preferred instruction with each app, and that the correlation coefficient between the two was 0.86. It was unclear to a reviewer exactly which data was being correlated, and which type of correlation coefficient was used.

We present the data that was correlated in Table 3.6. The Pearson product-moment correlation coefficient method was used to find a correlation coefficient of 0.86.

3.7.3 Statistical significance of the results

We make claims about the preferences and recommendations of the 'majority of the population' on page 87. To make these claims we conducted hypothesis tests from our sample data.

In Section 3.7.4, we present each individual hypothesis test conducted as part of this work in full detail.

In Section 3.7.5, we present additional hypothesis tests conducted using the binomial distribution. We have conducted this analysis because z-test was

Application	**Percentage of participants who preferred instruction with the application**	**Usability survey score**
Binary Search	84.78	81
Binary Numbers	70.05	68
CPU	69.17	63
Polynomial	67.36	69
Quicksort	58.31	51
Dijkstra	62.98	69

Table 3.6: Percentage of participants who preferred instruction with the application vs. usability survey score

used in the original hypothesis tests and it uses the normal distribution, an approximation of the binomial distribution.

These hypotheses were not suggested by the data. The primary purpose of the experiment was to investigate whether integrating educational software incorporating games and/or game design elements into the instruction of introductory computer science could increase student satisfaction and engagement relative to traditional instructional methods.

3.7.4 Z-test hypothesis test details

In this paper we used the z-test to make claims regarding the statistical significance of the results. The usage of the z-test was based on a belief that for $n = 36$, the z-test would be sufficient.

We re-evaluate the statistical analysis and claims made in this work below, and show all the work done in each hypothesis test.

Binary search application preference

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population prefers instruction with the binary search application.

Our survey indicated that 56/69 participants (81.20%) preferred the binary search application. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.
Note: the chosen significance level implies a critical value of 2.575.
Computing the test statistic:

$$z = (0.8120 - 0.5)/\sqrt{0.5 * (1 - 0.5)/69} = 5.18$$

Applying the decision rule: $z = 5.18 > 2.575$, therefore we reject the null hypothesis.

Binary numbers application preference

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population prefers instruction with the binary numbers application.

Our survey indicated that 42/62 participants (67.74%) preferred the binary numbers application. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.
Note: the chosen significance level implies a critical value of 2.575.
Computing the test statistic:

$$z = (0.6774 - 0.5)/\sqrt{0.5 * (1 - 0.5)/62} = 2.79$$

Applying the decision rule: $z = 2.79 > 2.575$, therefore we reject the null hypothesis.

Polynomial application preference

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population prefers instruction with the polynomial application.

Our survey indicated that 41/59 participants (69.49%) preferred the polynomial application. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.
Note: the chosen significance level implies a critical value of 2.575.
Computing the test statistic:

$$z = (0.6949 - 0.5)/\sqrt{0.5 * (1 - 0.5)/59} = 2.99$$

Applying the decision rule: $z = 2.99 > 2.575$, therefore we reject the null hypothesis.

Dijkstra application preference

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population prefers instruction with the Dijkstra application.

Our survey indicated that 34/49 participants (69.34%) preferred the Dijkstra application. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.
Note: the chosen significance level implies a critical value of 2.575.
Computing the test statistic:

$$z = (0.6934 - 0.5)/\sqrt{0.5 * (1 - 0.5)/49} = 2.71$$

Applying the decision rule: $z = 2.71 > 2.575$, therefore we reject the null hypothesis.

CPU application preference

On page 87 we stated that we could say with statistical significance (p-value 0.05) that the majority of the population prefers instruction with the CPU application.

Our survey indicated that 38/60 participants (63.33%) preferred the CPU application. We conduct the following hypothesis test at significance level $\alpha = 0.05$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-1.96 <= z <= 1.96$.
Note: the chosen significance level implies a critical value of 1.96.
Computing the test statistic:

$$z = (0.6333 - 0.5)/\sqrt{0.5 * (1 - 0.5)/60} = 2.07$$

Applying the decision rule: $z = 2.07 > 1.96$, therefore we reject the null hypothesis.

Binary numbers combination recommendation

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population recommends a combination of the binary numbers application and traditional instruction be used in the future.

Our survey indicated that 52/61 participants (85.25%) recommend a combination of the binary numbers application and traditional instruction be

used in the future. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.
Note: the chosen significance level implies a critical value of 2.575.
Computing the test statistic:

$$z = (0.8525 - 0.5)/\sqrt{0.5 * (1 - 0.5)/61} = 5.51$$

Applying the decision rule: $z = 5.51 > 2.575$, therefore we reject the null hypothesis.

CPU combination recommendation

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population recommends a combination of the CPU application and traditional instruction be used in the future.

Our survey indicated that 48/60 participants (80.00%) recommend a combination of the CPU application and traditional instruction be used in the future. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.
Note: the chosen significance level implies a critical value of 2.575.
Computing the test statistic:

$$z = (0.80 - 0.5)/\sqrt{0.5 * (1 - 0.5)/60} = 4.65$$

Applying the decision rule: $z = 4.65 > 2.575$, therefore we reject the null hypothesis.

Polynomial combination recommendation

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population recommends a combination of the polynomial application and traditional instruction be used in the future.

Our survey indicated that 48/58 participants (82.76%) recommend a combination of the polynomial application and traditional instruction be used in the future. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.
Note: the chosen significance level implies a critical value of 2.575.
Computing the test statistic:

$$z = (0.8276 - 0.5)/\sqrt{0.5 * (1 - 0.5)/58} = 4.99$$

Applying the decision rule: $z = 4.99 > 2.575$, therefore we reject the null hypothesis.

Quicksort combination recommendation

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population recommends a combination of the quicksort application and traditional instruction be used in the future.

We were unable to verify this claim at p-value 0.01. We are unsure why this calculation was originally made incorrectly. Whatever the case, we conduct a new hypothesis test (p-value 0.02)

Our survey indicated that 32/47 participants (68.09%) recommend a combination of the quicksort application and traditional instruction be used in the future. We conduct the following hypothesis test at significance level $\alpha = 0.02$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.3267 <= z <= 2.3267$.

Note: the chosen significance level implies a critical value of 2.3267.

Computing the test statistic:

$$z = (0.6809 - 0.5)/\sqrt{0.5 * (1 - 0.5)/47} = 2.48$$

Applying the decision rule: $z = 2.48 > 2.3267$, therefore we reject the null hypothesis.

Dijkstra combination recommendation

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population recommends a combination of the polynomial application and traditional instruction be used in the future.

Our survey indicated that 40/49 participants (81.63%) recommend a combination of the Dijkstra application and traditional instruction be used in the future. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$z = (\hat{p} - p_0)/\sqrt{p_0 * (1 - p_0)/n}$$

where $\hat{p}$ is the sample proportion, p_0 is the null hypothesis population proportion, p_A is the alternative hypothesis sample proportion, and n is the sample size.

Decision rule: Accept the null hypothesis if $-2.575 <= z <= 2.575$.

Note: the chosen significance level implies a critical value of 2.575.

Computing the test statistic:

$$z = (0.8163 - 0.5)/\sqrt{0.5 * (1 - 0.5)/49} = 4.43$$

Applying the decision rule: $z = 4.43 > 2.575$, therefore we reject the null hypothesis.

3.7.5 Binomial distribution hypothesis test details

For large values of n it can be argued that the shape of the binomial distribution is close enough to normal that the normal distribution can be used as an approximation.

We have found various guidelines for when a binomial distribution is close enough to normal that a normal distribution can be used as an approximation. For example, $np >= 10$ and $n(p-1) >= 10$ has been suggested as appropriate conditions for using a z-test, or that $np >= 5$ and $n(p-1) >= 5$ are appropriate conditions, or that $n > 30$ is an appropriate condition.

Given the subjectivity as to what value of n is large "enough", and the fact that using a z-test at all entails an approximation, we conduct additional hypothesis testing using the binomial distribution for the sake of providing additional analysis.

Binary search application preference

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population prefers instruction with the binary search application.

Our survey indicated that 56/69 participants (81.20%) preferred the binary search application. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who preferred the binary search application, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.01$.

Computing the test statistic:

$$pval = 2 * 8.41 \times 10^{-8} = 1.68 \times 10^{-7}$$

Applying the decision rule: $pval = 1.68 \times 10^{-7} < 0.01$, therefore we reject the null hypothesis.

Binary numbers application preference

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population prefers instruction with the binary numbers application.

Our survey indicated that 42/62 participants (67.74%) preferred the binary numbers application. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who preferred the binary numbers application, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.01$.
Computing the test statistic:

$$pval = 2 * 0.00357 = 0.00714$$

Applying the decision rule: $pval = 0.00714 < 0.01$, therefore we reject the null hypothesis.

Polynomial application preference

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population prefers instruction with the polynomial application.

Our survey indicated that 41/59 participants (69.49%) preferred the polynomial application. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who preferred the polynomial application, and Z is the number of participants who did not (note that $Z = n - Y$

where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.01$.

Computing the test statistic:

$$pval = 2 * 0.00189 = 0.00378$$

Applying the decision rule: $pval = 0.00378 < 0.01$, therefore we reject the null hypothesis.

Dijkstra application preference

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population prefers instruction with the Dijkstra application.

Our survey indicated that 34/49 participants (69.34%) preferred the Dijkstra application. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who preferred the Dijkstra application, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.01$.

Computing the test statistic:

$$pval = 2 * 0.00469 = 0.00938$$

Applying the decision rule: $pval = 0.00938 < 0.01$, therefore we reject the null hypothesis.

CPU application preference

On page 87 we stated that we could say with statistical significance (p-value 0.05) that the majority of the population prefers instruction with the CPU application.

The decision rule for this z-test was particularly close ($z = 2.07 > 1.96$), see section 3.7.4 for the full hypothesis test. We are unable to reject the null hypothesis using a binomial hypothesis test at significance level $\alpha = 0.05$, but are able to do so at significance level $\alpha = 0.06$.

Our survey indicated that 38/60 participants (63.33%) preferred the CPY application. We conduct the following hypothesis test at significance level $\alpha = 0.06$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who preferred the CPU application, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.06$.
Computing the test statistic:

$$pval = 2 * 0.0259 = 0.0518$$

Applying the decision rule: $pval = 0.0518 < 0.06$, therefore we reject the null hypothesis.

Binary numbers combination recommendation

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population recommends a combination of the binary numbers application and traditional instruction be used in the future.

Our survey indicated that 52/61 participants (85.25%) recommend a combination of the binary numbers application and traditional instruction be used in the future. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who recommend a combination of the binary numbers application and traditional instruction be used in the future,

and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.01$.

Computing the test statistic:

$$pval = 2 * 9.014 \times 10^{-9} = 1.803 \times 10^{-8}$$

Applying the decision rule: $pval = 1.803 \times 10^{-8} < 0.01$, therefore we reject the null hypothesis.

CPU combination recommendation

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population recommends a combination of the CPU application and traditional instruction be used in the future.

Our survey indicated that 48/60 participants (80.00%) recommend a combination of the CPU application and traditional instruction be used in the future. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who recommend a combination of the CPU application and traditional instruction be used in the future, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.01$.

Computing the test statistic:

$$pval = 2 * 1.59 \times 10^{-6} = 3.18 \times 10^{-6}$$

Applying the decision rule: $pval = 3.18 \times 10^{-6} < 0.01$, therefore we reject the null hypothesis.

Polynomial combination recommendation

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population recommends a combination of the polynomial application and traditional instruction be used in the future.

Our survey indicated that 48/58 participants (82.76%) recommend a combination of the polynomial application and traditional instruction be used in the future. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who recommend a combination of the polynomial application and traditional instruction be used in the future, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.01$.
Computing the test statistic:

$$pval = 2 * 2.26 \times 10^{-7} = 4.51 \times 10^{-7}$$

Applying the decision rule: $pval = 4.51 \times 10^{-7} < 0.01$, therefore we reject the null hypothesis.

Quicksort combination recommendation

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population recommends a combination of the quicksort application and traditional instruction be used in the future.

Our survey indicated that 32/47 participants (68.09%) recommend a combination of the quicksort application and traditional instruction be used in the future. We conduct the following hypothesis test at significance level $\alpha = 0.02$.

Null hypothesis: $p_0 = 0.5$
Alternative hypothesis: $p_A \neq 0.5$
Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who recommend a combination of the quicksort application and traditional instruction be used in the future, and Z is the number of participants who did not (note that $Z = n - Y$ where n is

the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.02$.

Computing the test statistic:

$$pval = 2 * 0.00931 = 0.0186$$

Applying the decision rule: $pval = 0.0186 < 0.02$, therefore we reject the null hypothesis.

Dijkstra combination recommendation

On page 87 we stated that we could say with statistical significance (p-value 0.01) that the majority of the population recommends a combination of the polynomial application and traditional instruction be used in the future.

Our survey indicated that 40/49 participants (81.63%) recommend a combination of the Dijkstra application and traditional instruction be used in the future. We conduct the following hypothesis test at significance level $\alpha = 0.01$.

Null hypothesis: $p_0 = 0.5$

Alternative hypothesis: $p_A \neq 0.5$

Test statistic:

$$pval = 2 * P(X >= Y) = 2 * P(X <= Z)$$

where Y is the number of participants who recommend a combination of the Dijkstra application and traditional instruction be used in the future, and Z is the number of participants who did not (note that $Z = n - Y$ where n is the number of participants). Note that we can use such a test statistic because the binomial distribution is symmetrical when $p = 0.5$.

Decision rule: Accept the null hypothesis if $pval >= 0.01$.

Computing the test statistic:

$$pval = 2 * 4.63 \times 10^{-6} = 9.26 \times 10^{-6}$$

Applying the decision rule: $pval = 9.26 \times 10^{-6} < 0.01$, therefore we reject the null hypothesis.

3.7.6 ANOVA hypothesis tests for quiz results

Our external examiner suggested analyzing the quiz results using analysis of variance testing. As a result, we perform ANOVA hypothesis tests for each application below.

Binary search

We conduct the following analysis of variance hypothesis test at significance level $\alpha = 0.05$. 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; $\mu_{ipad} = \mu_{trad}$
   H_A; $\mu_{ipad} \neq \mu_{trad}$

2. H_0; $\mu_{q1} = \mu_{q2} = \mu_{q3}$
   H_A; $\mu_{q1} \neq \mu_{q2} \neq \mu_{q3}$

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 2×3 mixed-design analysis of variance model where the between-subjects variable is what instructional method was used first (either iPad-first or traditional-first) and the within-subjects variable is the quiz scores (quiz 1, quiz 2, or quiz 3).

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 3.7, where SS is the sum of squares, df is the degrees of freedom, MS is the mean square, and F is the test statistic.

Applying the decision rules:

1. $F_{BS} = 4.75 > 4.05$, therefore we **reject the null hypothesis**.

2. $F_{WS} = 17.15 > 4.05$, therefore we **reject the null hypothesis**.

Source	**SS**	**df**	**MS**	**F**
Between Subjects	29.03	69		
Factor$_{\mathbf{BS}}$	1.9	1	1.9	4.75
Error	27.13	68	0.4	
Within Subjects	22.67	140		
Factor$_{\mathbf{WS}}$	4.47	2	2.23	17.15
Factor$_{\mathbf{BSxWS}}$	0.07	2	0.03	0.23
Error	18.13	136	0.13	
Total	51.7	209		

Table 3.7: Binary Search ANOVA Summary Table

3. $F_{BS\times WS} = 0.23 < 4.05$, therefore we fail to reject the null hypothesis.

The statistically significant between-subjects effect provides evidence that there was a difference in capability with the binary search concept between the group that used the iPad first and the group that experienced traditional methods first.

The statistically significant within-subjects effect provides evidence that the instruction provided to the students (whether using the iPad first or using traditional methods first) was effective.

The lack of statistically significant interaction does not allow us to suggest whether using the iPad first or traditional methods first is advisable.

Binary numbers

We conduct the following analysis of variance hypothesis test at significance level $\alpha = 0.05$. 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; $\mu_{ipad} = \mu_{trad}$
   H_A; $\mu_{ipad} \neq \mu_{trad}$
2. H_0; $\mu_{q1} = \mu_{q2} = \mu_{q3}$
   H_A; $\mu_{q1} \neq \mu_{q2} \neq \mu_{q3}$
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 2×3 mixed-design analysis of variance model where the between-subjects variable is what instructional method was used first (either iPad-first or traditional-first) and the within-subjects variable is the quiz scores (quiz 1, quiz 2, or quiz 3).

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 3.8, 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	30.87	63		
Factor$_{\mathbf{BS}}$	0.29	1	0.29	0.59
Error	30.58	62	0.49	
Within Subjects	13.33	128		
Factor$_{\mathbf{WS}}$	2.84	2	1.42	17.75
Factor$_{\mathbf{BSxWS}}$	0.18	2	0.09	1.13
Error	10.31	124	0.08	
Total	44.2	191		

Table 3.8: Binary Numbers ANOVA Summary Table

Applying the decision rules:

1. $F_{BS} = 0.59 < 4.05$, therefore we fail to reject the null hypothesis.
2. $F_{WS} = 17.75 > 4.05$, therefore we **reject the null hypothesis**.
3. $F_{BS \times WS} = 1.13 < 4.05$, therefore we fail to reject the null hypothesis.

The lack of statistically significant between-subjects effect does not allow us to claim whether there was a difference in capability with the binary numbers concept between the group that used the iPad first and the group that experienced traditional methods first.

The statistically significant within-subjects effect provides evidence that the instruction provided to the students (whether using the iPad first or using traditional methods first) was effective.

The lack of statistically significant interaction does not allow us to suggest whether using the iPad first or traditional methods first is advisable.

CPU

We conduct the following analysis of variance hypothesis test at significance level $\alpha = 0.05$. 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; $\mu_{ipad} = \mu_{trad}$
   H_A; $\mu_{ipad} \neq \mu_{trad}$
2. H_0; $\mu_{q1} = \mu_{q2} = \mu_{q3}$
   H_A; $\mu_{q1} \neq \mu_{q2} \neq \mu_{q3}$
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 2×3 mixed-design analysis of variance model where the between-subjects variable is what instructional method was used first (either iPad-first or traditional-first) and the within-subjects variable is the quiz scores (quiz 1, quiz 2, or quiz 3).

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 3.9, 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	14.31	59		
Factor$_{\mathbf{BS}}$	0.77	1	0.77	3.35
Error	13.54	58	0.23	
Within Subjects	28	129		
Factor$_{\mathbf{WS}}$	14.14	2	7.07	64.27
Factor$_{\mathbf{BSxWS}}$	0.72	2	0.36	3.27
Error	13.14	116	0.11	
Total	42.31	179		

Table 3.9: CPU ANOVA Summary Table

Applying the decision rules:

1. $F_{BS} = 3.35 < 4.05$, therefore we fail to reject the null hypothesis.
2. $F_{WS} = 64.27 > 4.05$, therefore we **reject the null hypothesis**.
3. $F_{BS \times WS} = 3.27 < 4.05$, therefore we fail to reject the null hypothesis.

The lack of statistically significant between-subjects effect does not allow us to claim whether there was a difference in capability with the CPU concept between the group that used the iPad first and the group that experienced traditional methods first.

The statistically significant within-subjects effect provides evidence that the instruction provided to the students (whether using the iPad first or using traditional methods first) was effective.

The lack of statistically significant interaction does not allow us to suggest whether using the iPad first or traditional methods first is advisable.

Polynomial

We conduct the following analysis of variance hypothesis test at significance level $\alpha = 0.05$. 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; $\mu_{ipad} = \mu_{trad}$
   H_A; $\mu_{ipad} \neq \mu_{trad}$

2. H_0; $\mu_{q1} = \mu_{q2} = \mu_{q3}$
   H_A; $\mu_{q1} \neq \mu_{q2} \neq \mu_{q3}$

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 2×3 mixed-design analysis of variance model where the between-subjects variable is what instructional method was used first (either iPad-first or traditional-first) and the within-subjects variable is the quiz scores (quiz 1, quiz 2, or quiz 3).

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 3.10, 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	13.66	58		
Factor$_{\mathbf{BS}}$	0.27	1	0.27	1.17
Error	13.39	57	0.23	
Within Subjects	25.33	118		
Factor$_{\mathbf{WS}}$	11.67	2	5.84	48.67
Factor$_{\mathbf{BSxWS}}$	0.27	2	0.13	1.08
Error	13.39	114	0.12	
Total	38.99	176		

Table 3.10: Polynomial ANOVA Summary Table

Applying the decision rules:

1. $F_{BS} = 1.17 < 4.05$, therefore we fail to reject the null hypothesis.
2. $F_{WS} = 48.67 > 4.05$, therefore we **reject the null hypothesis**.
3. $F_{BS \times WS} = 1.08 < 4.05$, therefore we fail to reject the null hypothesis.

The lack of statistically significant between-subjects effect does not allow us to claim whether there was a difference in capability with the polynomial concept between the group that used the iPad first and the group that experienced traditional methods first.

The statistically significant within-subjects effect provides evidence that the instruction provided to the students (whether using the iPad first or using traditional methods first) was effective.

The lack of statistically significant interaction does not allow us to suggest whether using the iPad first or traditional methods first is advisable.

Quicksort

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

Null and alternative hypotheses:

1. $H_0;\ \mu_{ipad} = \mu_{trad}$
   $H_A;\ \mu_{ipad} \neq \mu_{trad}$
2. $H_0;\ \mu_{q1} = \mu_{q2} = \mu_{q3}$
   $H_A;\ \mu_{q1} \neq \mu_{q2} \neq \mu_{q3}$
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 2×3 mixed-design analysis of variance model where the between-subjects variable is what instructional method was used first (either iPad-first or traditional-first) and the within-subjects variable is the quiz scores (quiz 1, quiz 2, or quiz 3).

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 3.11, 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	12.99	48		
Factor$_{\mathbf{BS}}$	2.08	1	2.08	9.04
Error	10.91	47	0.23	
Within Subjects	25.33	98		
Factor$_{\mathbf{WS}}$	8.34	2	4.17	37.91
Factor$_{\mathbf{BSxWS}}$	1.22	2	0.61	5.55
Error	10.44	94	0.11	
Total	32.99	146		

Table 3.11: Quicksort ANOVA Summary Table

Applying the decision rules:

1. $F_{BS} = 9.04 > 4.05$, therefore we **reject the null hypothesis**.

2. $F_{WS} = 37.91 > 4.05$, therefore we **reject the null hypothesis**.

3. $F_{BS \times WS} = 5.55 > 4.05$, therefore we **reject the null hypothesis**.

The statistically significant between-subjects effect provides evidence that there was a difference in capability with the quicksort concept between the group that used the iPad first and the group that experienced traditional methods first.

The statistically significant within-subjects effect provides evidence that the instruction provided to the students (whether using the iPad first or using traditional methods first) was effective.

The statistically significant interaction provides evidence that using the traditional instruction method first would lead to higher quiz scores.

Dijkstra

We conduct the following analysis of variance hypothesis test at significance level $\alpha = 0.05$. 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; $\mu_{ipad} = \mu_{trad}$
   H_A; $\mu_{ipad} \neq \mu_{trad}$

2. H_0; $\mu_{q1} = \mu_{q2} = \mu_{q3}$
   H_A; $\mu_{q1} \neq \mu_{q2} \neq \mu_{q3}$

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 2×3 mixed-design analysis of variance model where the between-subjects variable is what instructional method was used first (either iPad-first or traditional-first) and the within-subjects variable is the quiz scores (quiz 1, quiz 2, or quiz 3).

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 3.12, where SS is the sum of squares, df is the degrees of freedom, MS is the mean square, and F is the test statistic.

Applying the decision rules:

1. $F_{BS} = 0.3 < 4.05$, therefore we fail to reject the null hypothesis.

2. $F_{WS} = 62.36 > 4.05$, therefore we **reject the null hypothesis**.

Source	**SS**	**df**	**MS**	**F**
Between Subjects	12.75	48		
Factor$_{BS}$	0.08	1	0.08	0.3
Error	12.67	47	0.27	
Within Subjects	24	98		
Factor$_{WS}$	13.73	2	6.86	62.36
Factor$_{BSxWS}$	0.19	2	0.09	0.82
Error	10.08	94	0.11	
Total	36.75	146		

Table 3.12: Dijkstra ANOVA Summary Table

3. $F_{BS \times WS} = 0.82 < 4.05$, therefore we fail to reject the null hypothesis.

The lack of statistically significant between-subjects effect does not allow us to claim whether there was a difference in capability with the Dijkstra concept between the group that used the iPad first and the group that experienced traditional methods first.

The statistically significant within-subjects effect provides evidence that the instruction provided to the students (whether using the iPad first or using traditional methods first) was effective.

The lack of statistically significant interaction does not allow us to suggest whether using the iPad first or traditional methods first is advisable.

Chapter 4

Literacy App Study

The following work was published in the journal of Entertainment Computing:

K. Browne, C. Anand, E. Gosse, Gamification and serious game approaches for adult literacy tablet software, Entertainment Computing 5.3 (2014) 135146. (doi:10.1016/j.entcom.2014.04.003)

The work has been printed in this thesis under license (Elsevier Article Sharing Policy).

4.1 Abstract

In this paper, we overview the design of tablet apps we designed and built to teach literacy to adults, and present the results and conclusions derived from experiments performed with target users. Low adult literacy is a significant problem with a high economic cost both for the individuals and for society. Programs created to address low adult literacy face access and engagement barriers that tablet software may be able to help overcome. We designed three tablet apps, using two contrasting approaches of incorporating game-design elements to engage the users. We tested the apps with participants from the Brant Skills Centre, a non-profit organization that offers adult literacy programs in Brantford, Ontario. Though participants were divided on whether they preferred the apps to more traditional instruction, most participants preferred using the apps in addition to more traditional instruction. Based on this we conclude that gamification and serious game design approaches were effective at increasing learner engagement, and we propose a direction for future research.

4.2 Introduction

The high prevalence of low literacy skills in adults has serious negative consequences for society and the individuals affected. Literacy is frequently measured on a scale of 1 to 5, with level 3 being equivalent to high school completion and considered by experts as the minimum required for coping with the increasing demands of the information economy[DMT05]. The International Adult Literacy and Skill Survey found that 48% of the Canadian population over 16, representing 12 million people, have literacy skills below this level[BTNP05].

The employment, economic and health impacts of low literacy skills have been extensively documented by international and national organizations [Hig08, Lit01]. We highlight a few statistics here to illustrate the scope of the problem. People with low literacy skills are more likely to have lower rates of employment and to work in low-skill jobs; roughly 57% of Canadians age 16-65 at a level 1 literacy level were employed compared to about 70% at level 2 and more than 80% of those at levels 4-5[BTNP05]. In the USA, a 2005 report by the National Assessment of Adult Literacy found that 93 million people lack the literacy skills required to complete the education or job training required by current and future jobs[KGB05]. With employment growth polarizing into high-skill, high-wage jobs and low-skill, low-wage jobs due to automation of routine tasks by computer software and off-shoring of middle-skill jobs[A$^+$10], the employment prospects of low-literacy workers will likely only become worse in the future. In the USA, an estimated $106 to $238 billion dollars in annual health-care costs can be attributed to poor health knowledge and behaviour resulting from low literacy[VTRB07, YJM$^+$09].

The total potential enrollment and drop-out rates of programs designed to address low literacy are disappointing. Less than half of those who reach out to a literacy organization register for a program, and of those 30% will drop out[Lon01]. Under 10% of Canadians who could benefit from a literacy program enroll[Lit01], with job, money problems, lack of childcare and transportation cited as reasons.

Tablet computers have recently exploded in popularity, reaching 116 million in sales worldwide in 2012 and are projected to grow to 468 million by 2017[Gar13]. These devices provide an opportunity for novel and disruptive approaches to the problem of low adult literacy. Tablet software may be more cost effective than human instruction, alleviating the resource issue. A tablet can provide an interactive learning experience that doesn't require the learner to leave home, alleviating the lack of childcare and transportation issues. Tablet software may also provide a more fun and engaging method

of learning that could increase adult participation in literacy education and / or lower literacy-program drop-out rates. Tablet devices offer a direct-control user interface (touchscreen) that could be easier to use, particularly for low-literacy users, than traditional PC or laptop home computers with keyboards and indirect pointing devices (mice, touchpads). Finally, Megalingam et al. have proposed a tablet device called EduPad to reduce rural adult illiteracy in India[MRSD12], motivated by the idea that the device could make up for a lack of qualified personnel or adequate infrastructure.

In this paper we explore using iPad software to teach literacy skills to low-literacy adults, motivated by the above situation and possible opportunity. For this we sought out existing literature on educational iPad software design and user experience, and on using gamification as a design approach due to its claim of influencing user behavior and increasing engagement.

Kayne Toukonen looked at how the features of a tablet device could facilitate learning in the form of what he termed *dynamic electronic textbooks*[Tou11]. Toukonen pays particular attention to Brian Cambourne's eight Conditions of Learning[REZ03] that Cambourne felt necessary for the acquisition of language, connecting each of these eight conditions to the possible software features enabled by tablets. For example, it is suggested that the learning condition *engagement* can be facilitated by using virtual worlds in tablet applications.

There have been some studies analyzing the effectiveness of tablets for education. For example, Houghton Mifflin Harcourt conducted a year-long pilot study in a middle school in Riverside, California in which an iPad application was used to increase student performance by 20% in algebra compared to peers who used textbooks[Che12]. In work by Sarah Henderson and Jeff Yeow, a New Zealand primary school used iPads in classrooms with students aged 5-12, and they reported that the low learning curve, high portability, and instant boot-up time of the iPads were recognized as positive aspects of using the tablets as learning tools[HY12]. Henderson and Yeow's findings and that of others[CG11] focus less on the effectiveness of the design of any particular application but more so on the aggregate results from using iPads as a learning tool; for example, the organizational challenges that using the tablets present.

The authors have previously created iPad apps for secondary and introductory post-secondary computer science education with promising results. A common design approach in these apps was having the user *learn-by-doing*; the user would manipulate representations or simulations of a problem domain. We believe this design approach is best described as experiential learning. Though experiential learning has several definitions, properties and models[Gen90], for the purposes of our software we focus on the fundamental feature of learning

through the application of concepts in an interactive setting.

Gamification is another design approach which has produced promising increases in engagement in several different contexts. Gamification can be defined as the "usage of game design elements to motivate user behaviour in non-game contexts"[Det11]. A well-known example would be the location-based social network Foursquare that rewards users for checking in to their current location with points and allows the user to become the "mayor" of that location. Other examples include startup company ZamZee[Zam12] that was able to use a gamification-based app to increase physical activity in children by 60%, and the Greater Washington Give Day who were able to use gamification to drive 2 million dollars in fundraising in one day[Bar12].

While it's clear gamification has produced some promising results, the design approach does come with significant risk. One of the world's leading IT research and analysis companies, Gartner has warned that 80% of all gamification apps will fail to meet their objectives due to poor design[Gar12]. There has also been little academic research into the definition of gamification itself, and whether it constitutes a truly new and distinct phenomenon. One one can also distinguish between gamification and "serious games", that is a full-fledged game for non-entertainment purposes, but this boundary is subjective.

There have been a few studies looking at educational-tablet game software. Wattanatchariya et al. developed a game called "Drop Donuts" for teaching wind and gravity concepts and report on survey results from 12 university student participants[WCD11]. Feng Yan developed an iPad application "A Sunny Day" to educate children with autism, which features mini-game and puzzle-game elements, and tested it with autistic children and their parents[Yan11]. Yan's work is especially valuable because of attention to application-design elements, to testing of the interface with target users, and reporting on what did and didn't work and why. For example, their analysis suggests that the software should have had clearer objectives and provided rewards more quickly after completing tasks. The paper concludes that the app offers a cost-effective therapeutic approach relative to existing methods. We were not able to find similar work examining gamification approaches to adult-literacy tablet software and consider this a significant gap in the literature that demands exploration.

In this paper, we present the design and user experience experiment results for three iPad applications developed to teach literacy concepts to low-literacy adults. The apps incorporate game-design mechanics in different ways, and we hypothesize using these apps will improve the educational experience through increased motivation, engagement, and learning. The study was done

in conjunction with the Brant Skills Centre, a non-profit organization that provides programs for literacy and other essential skills for adults in Brantford, Ontario. In Section 4.3, we present the design of the applications and how they incorporate game mechanics. In Section 4.4, we discuss the design of the experiment that took place. In Section 4.5, we provide the the results of this study, and in Section 4.6 we conclude that learner engagement was improved and propose a direction for future research.

4.3 Application Design

Three apps were designed to teach literacy concepts to experiment participants from the Brant Skills Centre. Because the clients of the Brant Skills Centre were at varying levels of low literacy, we chose to focus on three areas that tended to be problems across these levels: homophones, punctuation (period, question mark and quotation placement), and comma placement. All of the apps incorporate some form of experiential learning by having the user *do* what the app is trying to teach them. The homophone app and punctuation app were developed with two contrasting approaches to gamification of educational software. The comma placement app was actually developed in response to experiments conducted with the homophone app, as explained in Section 4.5.

In a paper focusing on the definition of gamification, Deterding et al distinguish between gamification from serious games as follows, "what distinguishes gamification from regular entertainment games and serious games is that they are built with the intention of a system that includes elements from games, not a full game proper"[DDKN11]. The homophone app is intended as an educational app that includes elements of gamification to motivate user behavior, whereas the punctuation app is intended as a serious game where literacy-skill improvement is the primary goal rather than entertainment. The design of the homophone app was most influenced by Cambourne's conditions of learning, and the design of the punctuation app was most influenced by the concept of flow[Csi97].

4.3.1 Homophone App

The homophone app was built to teach the differences between the following homophones:

- It's / Its
- We'll / Will

- You're / Your
- Two / Too / To
- Principle / Principal
- They're / There / Their

Overview

The homophone app begins with a title screen, shown in Figure 4.1, which presents the user with six homophone suboptions, allowing the user to choose what they would like to learn. Each option has a grayed-out check mark next to it if the user has not successfully completed the practice option provided in the submenu.

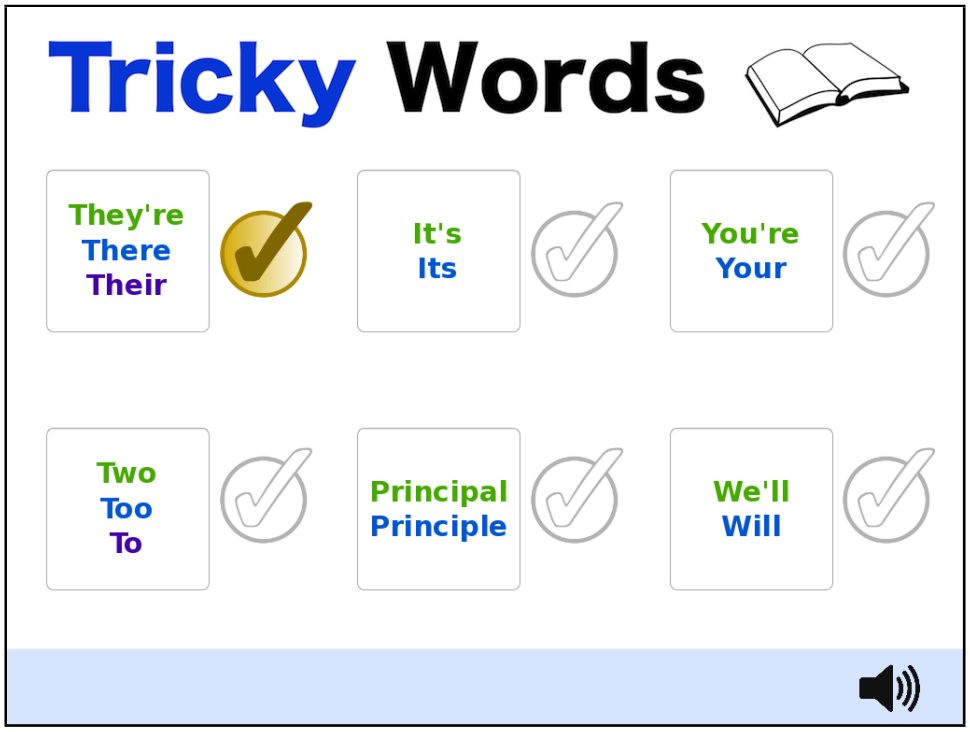

Figure 4.1: Title screen

As shown in Figure 4.2, each homophone submenu has the same three options: lesson, examples, and practice. The homophone submenu also contains a practice score, with five smaller check marks on the left-hand side and a large check mark on the right-hand side all initially grayed out. The users can

earn up to five green check marks on the left-hand side by doing the practice activity, and are awarded a gold check mark on the right-hand side (and in the title screen) when they have earned all five green check marks.

The lesson option brings the user to a new screen which gives a brief overview of the differences between the homophones containing a mixture of text and accompanying audio, illustrated with an example usage of each word.

Figure 4.2: Subscreen

The examples option shown in Figure 4.3 brings the user to a screen where they can tap a button for each homophone and are delivered an example usage of that homophone with text and audio. The user is delivered a new example each time they tap a button in the examples screen.

As shown in Figure 4.4, the practice option begins with a short non-interactive demonstration where a drawing of a hand drags a homophone into the blank space in a sentence and is awarded a green check mark for doing so. The practice option only does this the first time it is loaded so the user knows what is expected of them.

The user is given a series of five sentences with blank spaces for a homophone. Each sentence is presented both as text and audio. When the user drags the correct homophone into the blank space of the sentence, a chime

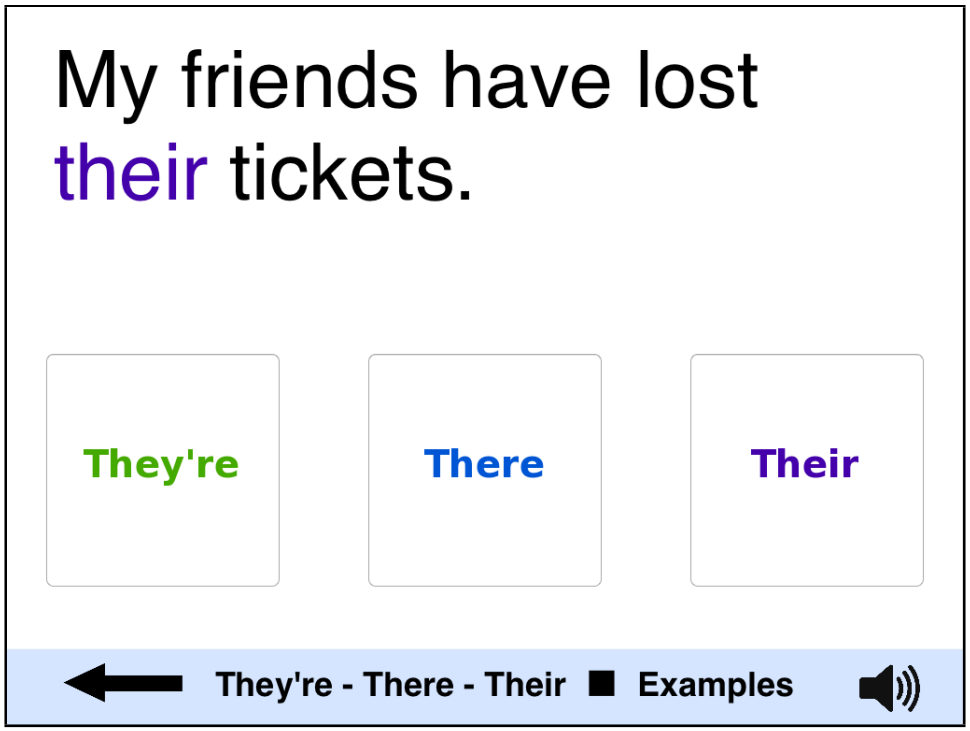

Figure 4.3: Examples

with an approving sound is played and a green check mark moves horizontally across the screen as shown in Figure 4.5.

When the user drags an incorrect homophone into the blank space of the sentence, a chime with a disapproving sound is played and a yellow circle with a diagonal line through it (i.e. a no symbol) moves horizontally across the screen as shown in Figure 4.6. The yellow no symbol was chosen because the usual red symbol would be too discouraging for participants. Though the meaning of sound effects are subjective and difficult to describe in writing, the disapproving sound for an incorrect answer was similarly sensitive in that it was more akin to a "missed shot" sound than the sort of "buzzing error" sound that you might hear on a TV game show.

The highest score out of five that the user has achieved is presented on the associated homophone submenu. The user can try the practice option as many times as they want. A back-arrow button in the bottom left allows users to move back to the previous menu at all times, and a speaker button at the

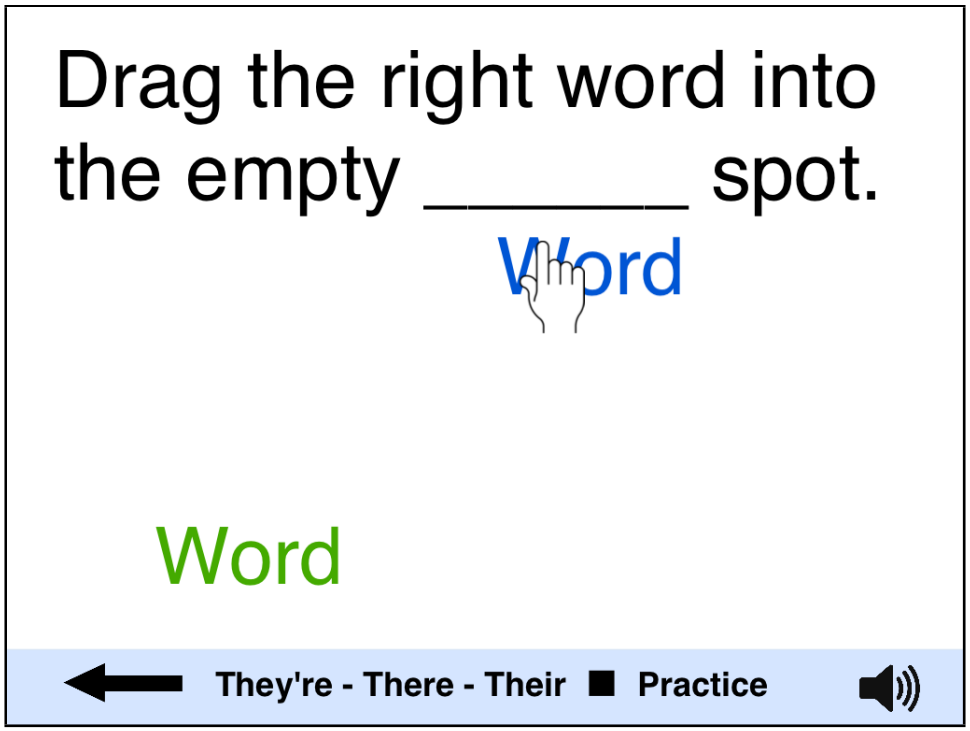

Figure 4.4: UI demonstration

bottom right allows the user to turn on or turn off the audio.

Design Approach

The overall design of the app is motivated by Cambourne's conditions of learning and Kayne's interpretations for how tablet software can fulfill these conditions[Tou11]. The conditions are summarized as follows:

1. **Immersion** - learners need to be immersed and constantly saturated in that which is to be learned.

2. **Demonstration** - learners need to receive many stimulating demonstrations of desired outcomes.

3. **Engagement** - learners must be engaged in the learning process while being immersed in the learning environment and viewing demonstrations.

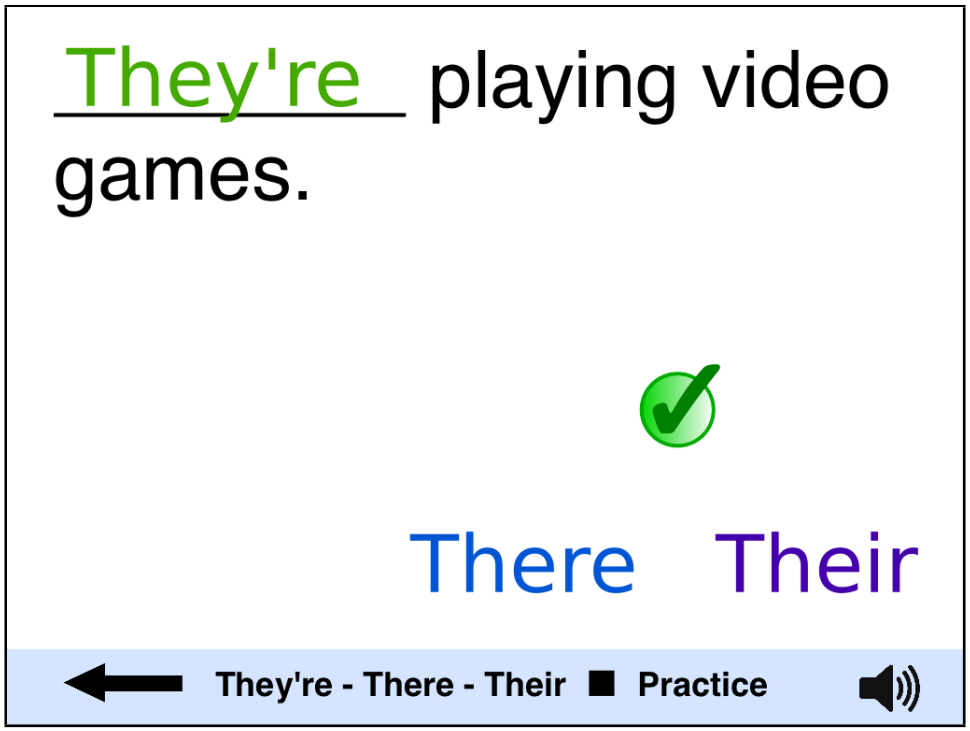

Figure 4.5: Right answer

4. **Expectations** - learners are influenced by expectations, which are powerful shapers of behavior.
5. **Responsibility** - learners need to make their own decisions about when, how, and what "bits" to learn.
6. **Employment** - learners need time and opportunity to use and practice new learning in realistic ways.
7. **Approximation** - learners must be free to approximate desired study, as mistakes are essential for learning to occur.
8. **Response** - learners must receive relevant, appropriate, timely, non-threatening feedback.

The app is designed to give users *responsibility* by allowing users to decide which homophones to learn (title screen) and how to learn them (through

Figure 4.6: Wrong answer

a lesson, examples and / or practice). No portion of the app has a time limit so users are free to learn when they want to learn as well. *Immersion* is facilitated by the audio that accompanies the text throughout the app. The examples feature of the app aims to fulfill the *demonstration* learning condition. *Expectations* are set out for the user by showing the grayed-out check marks when the app first launches and allowing them to earn green and gold check marks through successful practice. *Employment* of the concepts is provided by the practice feature of the app which allows users to practice their understanding in a realistic setting. The users are free to *approximate* their understanding as there is no penalty for making mistakes during the Practice portion of the app; users can redo the Practice portion as much as they like and are only rewarded with check marks for correct answers and do not lose anything for incorrect answers. These check marks and yellow no symbols that the user receives during practice facilitate the *response* learning condition.

Finally the *engagement* condition is facilitated by the gamification elements of the app:

- **Badges** - green and gold check marks as rewards for successful practice.
- **Levels** - each homophone submenu provides a discrete objective for the user.
- **Short, medium, and long term goals** - earning an individual green check mark is a short-term goal, earning a gold check mark for a homophone submenu is a medium-term goal, and earning all of the gold check marks is a long-term goal.

Distinguishing between short term goals with a green check mark and medium term goals with a gold check mark creates a visual layering of goals that take increasing amounts of time to complete but that come with increasing rewards is a key feature of many games[DM12].

4.3.2 Punctuation App

The punctuation app was built to teach when to use a period at the end of a sentence, when to use a question mark, and where quotation marks should be used (around phrases to show what words are being said, as well as around titles of stories, songs, books and movies).

Overview

The punctuation app does not begin with any sort of menu, but begins immediately with an introductory demonstration where a period is dragged by a drawing of a hand to the end of a sentence. From this screen shown in Figure 4.7, the user is then left to drag the remaining period to the end of the next sentence. A horizontal bar at the top of the screen next to a drawing of a clock tells the user how long they have to drag the period to the end of the sentence; the bar will turn yellow when only 25% of the time alloted is remaining.

If the user completes this step, then as seen in Figure 4.8, they will begin a virtual "tour of Italy" that simultaneously attempts to teach the user the correct usage of periods, question marks, and quotation marks. Each screen presents the user with a series of sentences which require the appropriate punctuation marks to be dragged into the correct places before the timer runs out in order to move on to the next screen. The user receives an explanation about how to use each punctuation mark correctly in the first screen that

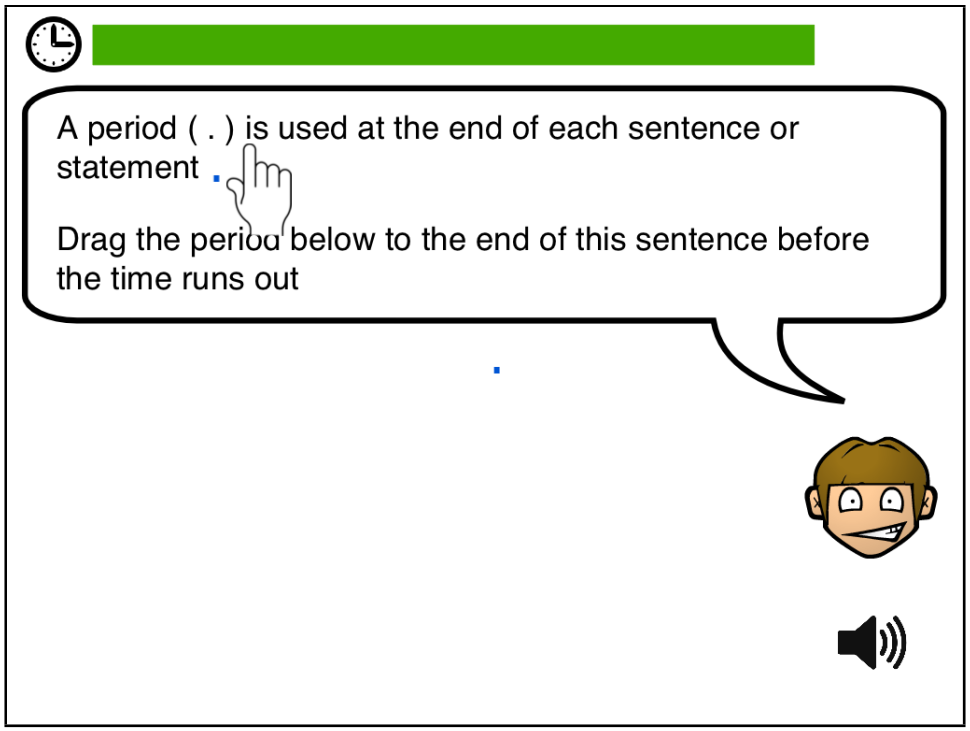

Figure 4.7: Punctuation app introduction

the punctuation mark is used. If the user drags a punctuation mark into the correct position, an approving sound is played. If the user drags a punctuation mark into an incorrect position, it moves back to its original position below the text. If the user does not complete the task in time, the correct answer is presented to the user as an audio explanation as to why this answer is correct. The user then repeats the same screen until they drag the correct punctuation marks into the correct positions.

The screens of the punctuation app get increasingly difficult over time, as at first the user is given only one or two punctuation marks to drag into the correct position before being given three and then finally four as shown in Figure 4.9.

There are ten screens in the punctuation app. All of the text in each screen is also presented to the user as audio. When the user completes the final screen, they are given a congratulatory goodbye by the character who has guided them through the tour in the lower right-hand side of the screen.

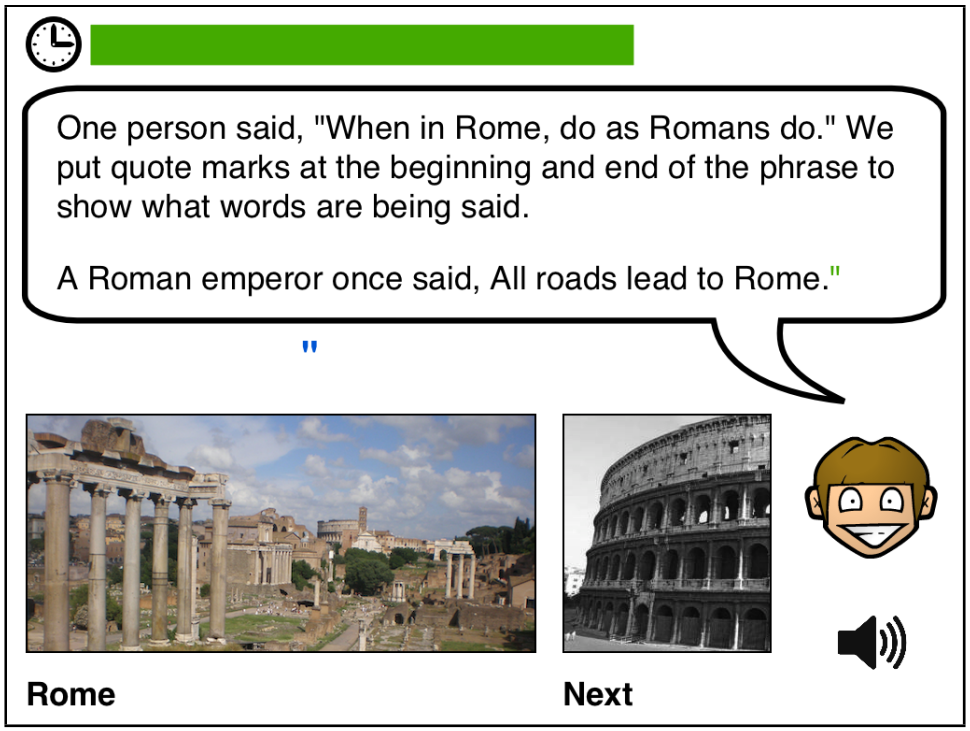

Figure 4.8: Punctuation app tour

The images of Italy that appear throughout the app were intended to provide an aesthetically pleasing reward for reaching each screen, with a grayed-out image of the next screen to provide some motivation to see what's next.

Design Approach

In contrast to the "learning system with game design elements" approach of the homophone app where the user self-consciously directs what, how, and when they learn, the punctuation app was designed to have the users experience *flow* and a loss of reflective self-consciousness as they become absorbed in playing the game. Nakamura and Csíkszentmihályi[NC09] describe flow as a subjective experience that seamlessly unfolds from moment to moment with the following features:

- intense and focused concentration on the present moment;
- merging of action and awareness;

Figure 4.9: Increasing difficulty

- loss of reflective self-consciousness;
- a feeling of control over the activity;
- distortion of temporal experience (i.e. that time has passed faster than normal);
- experience of the activity as intrinsically rewarding.

They also define the conditions for entering a state of flow to include the perception of challenges that stretch but do not exceed existing skills, as well as clear goals and immediate feedback about progress being made. As adapted from Csíkszentmihályi[Csi97], the earlier model of flow emphasized a balance between perceived opportunities and skills as shown in Figure 4.10. The current model of flow shown in Figure 4.11 has 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).

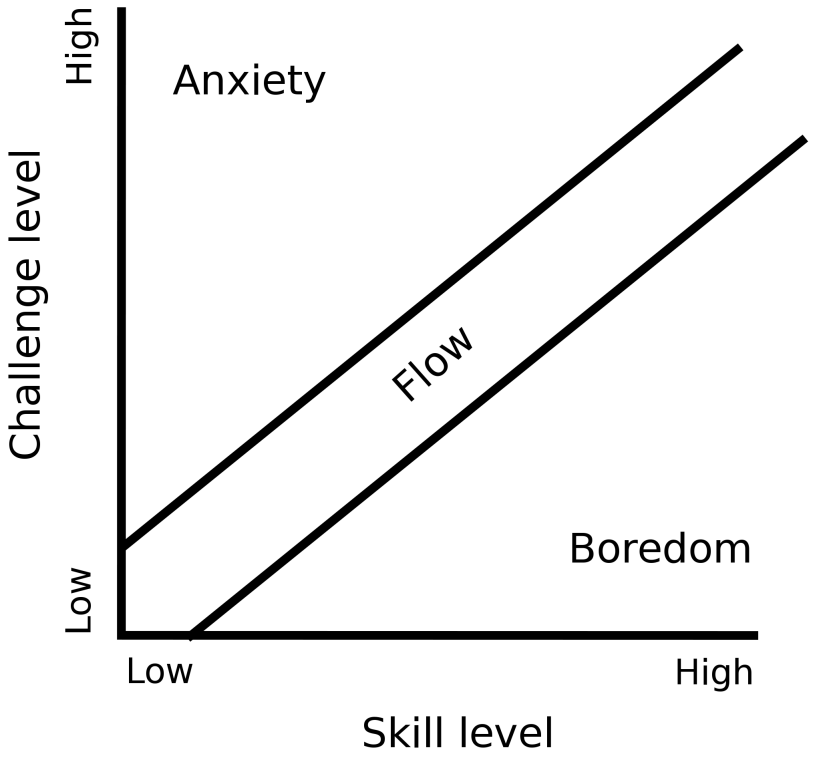

Figure 4.10: Earlier model of flow

The punctuation app was designed to induce a sense of flow through the clear goals and immediate feedback of moving the punctuation marks to the correct positions in the text. The timer at the top of the screen was meant to help focus the user's attention on the present moment as it would discourage them from losing focus on the task at hand. A tour of Italy was chosen as the narrative for the app because vacations and tours are intrinsically motivating activities.

Facilitating the perception that challenges stretch but do not exceed skills was done in two ways. First, any time a new concept was introduced, it is explained to the user (how the UI functions, the correct usage of periods, question marks, and quotation marks). This was done to ensure that the challenges being posed do not exceed the user's perceived skills. Second, the app gets progressively more difficult from one screen to the next by introducing

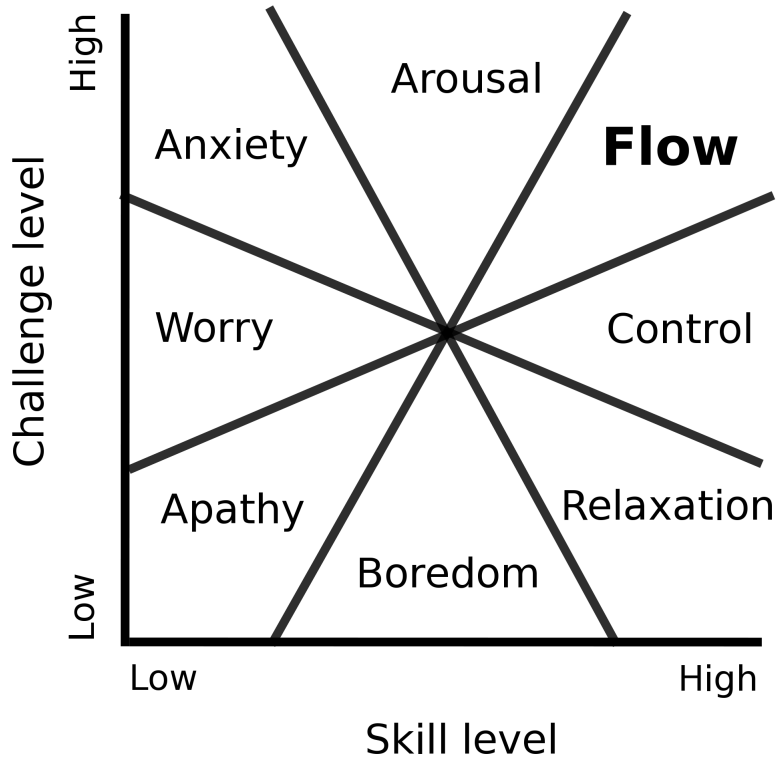

Figure 4.11: Csíkszentmihályi model of flow

more difficult problems (more text, more punctuation marks to place on the text). This was done to progressively stretch and ideally expand the user's skills, like climbing a set of stairs to a better understanding of the concepts. While more advanced users could become apathetic when faced with the initial easy screens which do not stretch their skills, we try to mitigate this by allowing advanced users to quickly reach the more difficult screens.

Other game features besides flow incorporated into the design of the app include the *narrative* of the tour of Italy which involves connected events with a beginning and an end for the user to progress through. And just like stops on a tour bus, each stop has a maximum duration made visually apparent by time bar.

4.3.3 Comma App

The comma app resembles the practice feature of the punctuation app, but with a few differences. The app begins with a demonstration where an image of a hand drags a comma to the correct position in a sentence. The user is then given a cycle of five sentences where they must drag the comma to the correct position. If the user drags the comma to the correct position they receive an approving sound and a green check mark as shown in Figure 4.12.

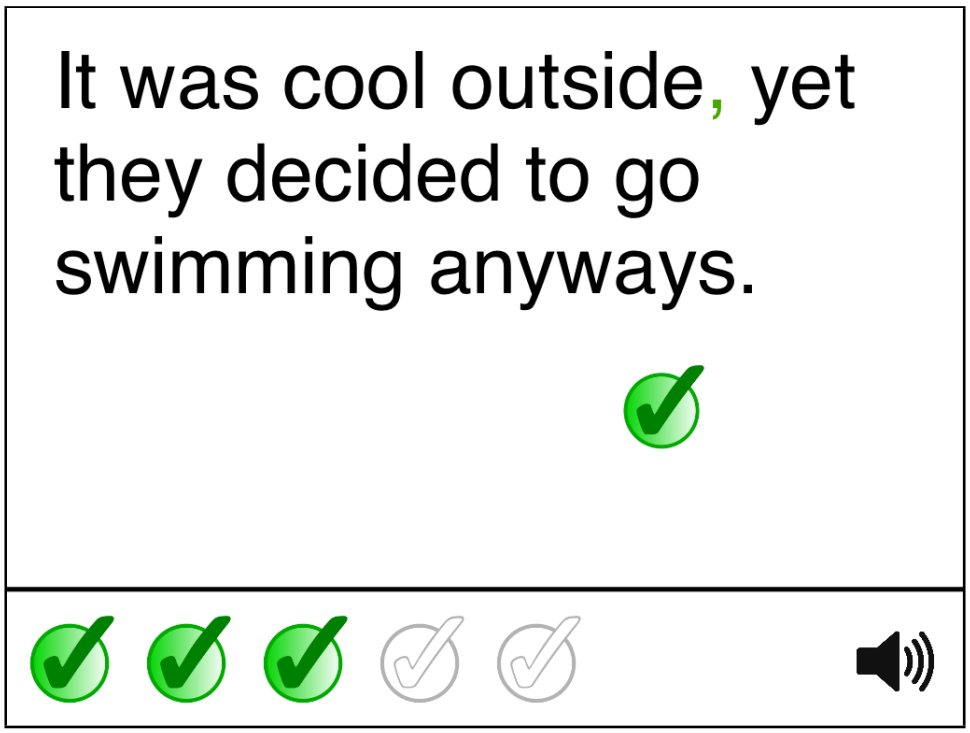

Figure 4.12: Comma app right answer

If the user drags the comma to an incorrect position, then, as shown in Figure 4.13, the correct answer is explained using audio. From here the user moves on to the next of the five sentences which has not been completed correctly.

The five sentences remain in the cycle until all of them have been completed correctly. A series of five grayed-out check marks in the bottom left-hand corner of the screen turn into green check marks as the user correctly completes each sentence. When the user completes all five sentences correctly, they are awarded a gold check mark and the app is complete.

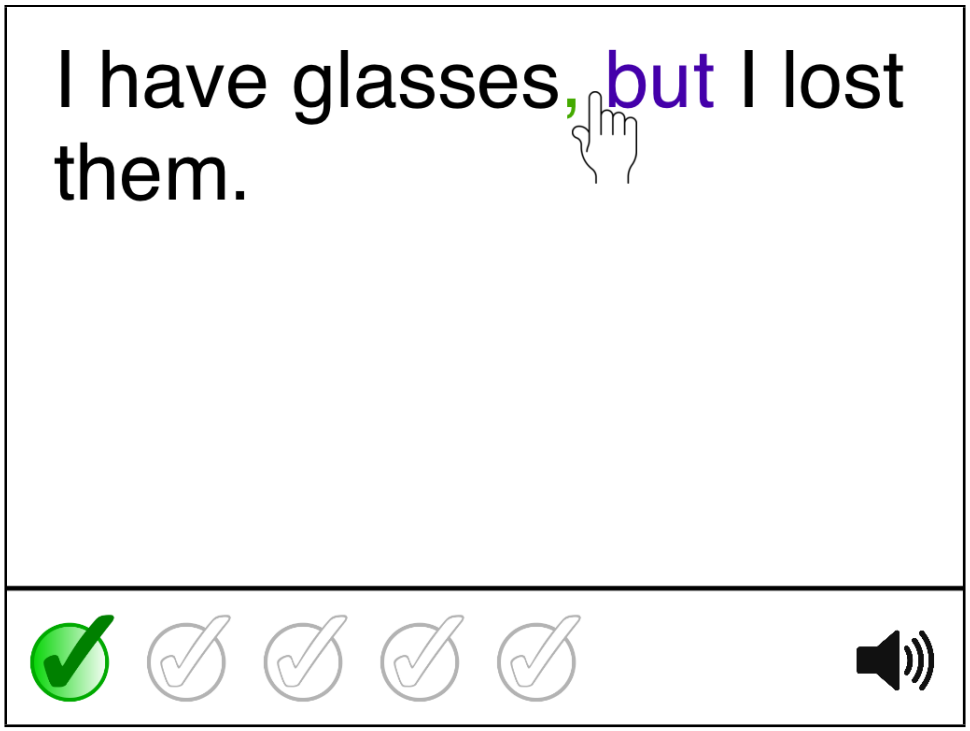

Figure 4.13: Correct answer demonstration

Design Approach

Experiments conducted with the homophone app inspired the creation and subsequent test of the comma app, which is essentially a modified version of the practice feature of the homophone app, and we explain the reception of the app in Section 4.5.

4.4 Experiment Design

Clients of the Brant Skills Centre were invited to participate in a study about teaching literacy with iPad applications. Staff spoke to them directly in conversations or during class. Clients who chose to participate were given consent forms to sign to confirm their participation and make them aware of the study details. We made sure that the consent forms for this study were written in plain language, and that if the client was unable to read the consent form

unassisted or had any questions, that the details were also be explained to them verbally to ensure that we had their consent.

The sessions took place in a classroom at the Brant Skills Centre and were conducted by Kevin Browne and Elizabeth Gosse. Participants received a $10 Tim Hortons gift card for each experiment session that they participated in to compensate them for their time and motivate their participation in the study. The study was approved by the McMaster Research Ethics Board.

4.4.1 Experiment Session Protocol

The following protocol was followed with each experiment participant. The protocol refers to the pre-experiment questionnaire in Section 4.4.4, the user experience survey in Section 3.4.3, and the post-experiment questionnaire in Section 4.4.6.

Four one-hour experiment sessions were conducted, two sessions in one week and two sessions in the following week. Participants could only attend one session per week, and some participants did not attend a session in both weeks. In the first week's session, the homophone app was tested, and in the second week's session, the punctuation app was tested followed by the comma app. The following procedure was used during each session for the homophone-app and punctuation-app experiments:

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 the first quiz sheet.
5. The participants were instructed using either traditional methods or the iPad app.
6. The participants completed the second quiz sheet, and the user experience survey if the iPad app had just been used.
7. The participants were instructed using either traditional methods or the iPad app.
8. The participants completed the third quiz sheet, and the user experience survey if the iPad app had just been used.

9. The participants completed the post-experiment questionnaire.

Each week, both presentation orders were used in one session (iPad app followed by traditional instruction versus traditional instruction followed by iPad app). The time spent on each method was roughly equivalent, about 15 or so minutes, or until the participants felt like going ahead and doing the next practice quiz.

Traditional instruction consisted of Brant Skills Centre instructor author3 teaching the concepts using a lecture and whiteboard, as well as pen-and-paper practice questions (different from the quiz sheets used in the above procedure). Author3's lecture method was highly interactive and tailored to the unique learning styles of the participants, as discussed in Section 4.5.

During the second week's session, after following the above procedure, we also did a brief experiment of the comma app. As is discussed further in Section 4.5, the comma app was developed quickly in response to the results of the first week's experiment that indicated a combination of traditional instruction and the iPad software may be most helpful. In the first session of the second week, the participants were given a brief lecture on comma placement, then practiced their understanding with pen and paper practice sheets, followed by another brief lecture on comma placement, followed by practice with the comma app. In the second session of the second week, the order of the lecture and accompanying practice method was reversed. In both instances, the participants completed the comma-experiment questionnaire found in Section 4.4.7.

The first week's sessions had a combined 14 participants (8 in the first session and 6 in the second), and the second week's sessions had a combined 13 participants (7 in the first session and 6 in the second).

4.4.2 Quantitative Observations

Quantitative observations were recorded using the quiz sheets to test the participants throughout the homophone-app and punctuation- app experiments. The quiz sheets consisted of one page of relevant questions. In the case of the homophone experiment, the participants were asked to fill in the blank word of a sentence with the correct word. In the case of the punctuation experiment, the participants were asked to fill in the missing punctuation of a sentence. The same quiz was given three times in an attempt to assess if learning was taking place, and how much, over the course of the experiments. The homophone quiz had a maximum score of 19, and the punctuation quiz had a maximum score of 10.

4.4.3 Qualitative Observations

After the session was over, a casual verbal discussion with the participants as a group and participants individually was used for further insights into the effectiveness of the software. Observations from these discussions were recorded in writing. Verbal expressions, reactions, and comments made by the participants during the study were also be recorded in writing as study data. Qualitative observations as to how users perceived each app were also recorded with the user experience survey in Section 4.4.5.

4.4.4 Pre-Experiment Questionnaire

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

- Gender (Male/Female)
- Age
- Handedness (Right/Left)

The participants were also asked to rate their reading ability from 1 to 5, if 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, if 1 is "not well at all" and 5 is "I can use the iPad perfectly well".

4.4.5 User Experience 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 learn _____.
- **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.
- **S9** I liked the app's graphics.
- **S10** I liked the app's voices / sound.

The participants could choose from: strongly disagree, disagree, neutral, agree, and strongly agree. For analysis purposes, these descriptions were assigned numeric values 1-5 from strongly disagree to strongly agree. An overall user experience score from 1-5 can also be computed by averaging the assigned numeric values. In the case of statement S5, the blank space was replaced with the concept being taught in the session. This and other wordings of the user experience survey statements were done under the advisement of Brant Skills Centre staff.

4.4.6 Post-Experiment Questionnaire

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

1. Did you prefer learning using the iPad app or learning from the instructor's lesson? (check one)
2. In the future should _____ be taught only using the iPad app, only using the instructor's lesson, or both? (check one)

The blank space was again replaced with the concept being taught in the session.

4.4.7 Comma-Experiment Questionnaire

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

1. I liked using the app to do practice exercises during the lesson.
2. In the future, I would prefer doing practice exercises during lessons using the iPad app, only using the instructor's lesson, or either? (check one)

The first question allowed participants to choose from: strongly disagree, disagree, neutral, agree, and strongly agree. Again for analysis purposes, these descriptions were assigned numeric values 1-5 from strongly disagree to strongly agree.

4.5 Results and Discussion

The first week of sessions where the homophone experiment took place had 14 participants total, 8 males and 6 females, 2 left-handed and 12 right-handed, with ages ranging from 21 to 63 and an average age of 41 (standard deviation of 13.5). The second week of sessions where the punctuation experiment and comma experiment took place had 13 participants total, 7 males and 6 females, 3 left-handed and 10 right-handed, with ages ranging from 21 to 54 and an average age of 42 (standard deviation of 11).

The participants' ratings of their reading ability and ability to use an iPad are presented in Figure 4.14. The participants' ratings of their ability to use an iPad are noticeably higher in the second week of sessions, which may indicate increased confidence in using the iPads after doing so during the first week's sessions.

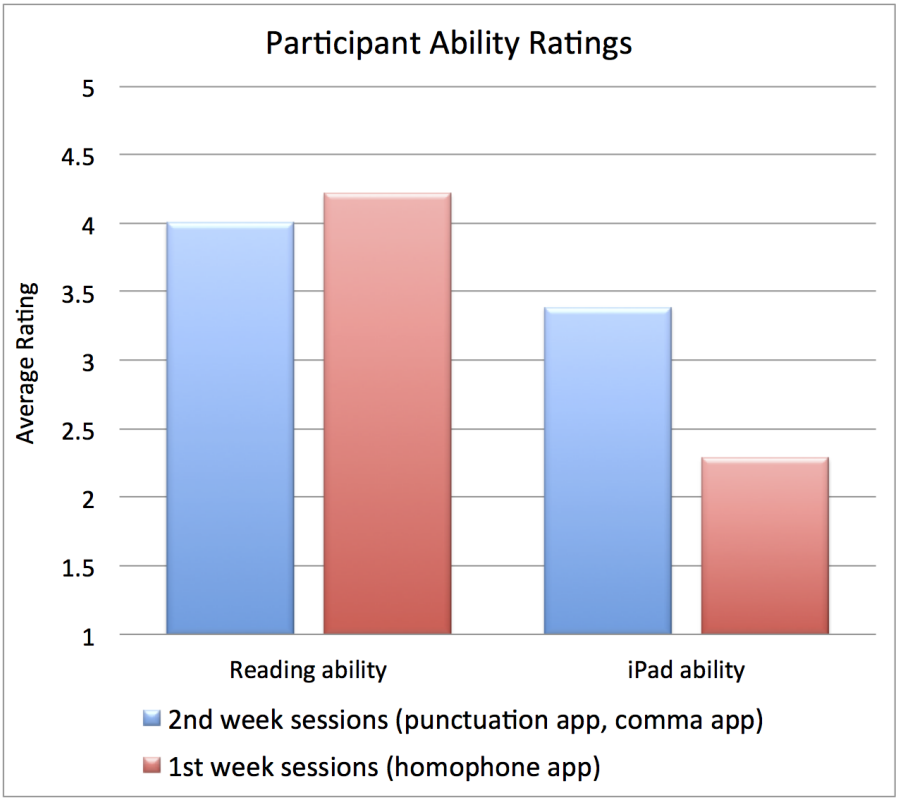

Figure 4.14: Participant ability ratings

4.5.1 Homophone Experiment

The homophone app received overall positive user experience feedback as shown in the survey results in Figure 4.15. The lesson and example features of the app were used by the participants and by all accounts were well received, but it was definitely the practice feature of the app that participants used most.

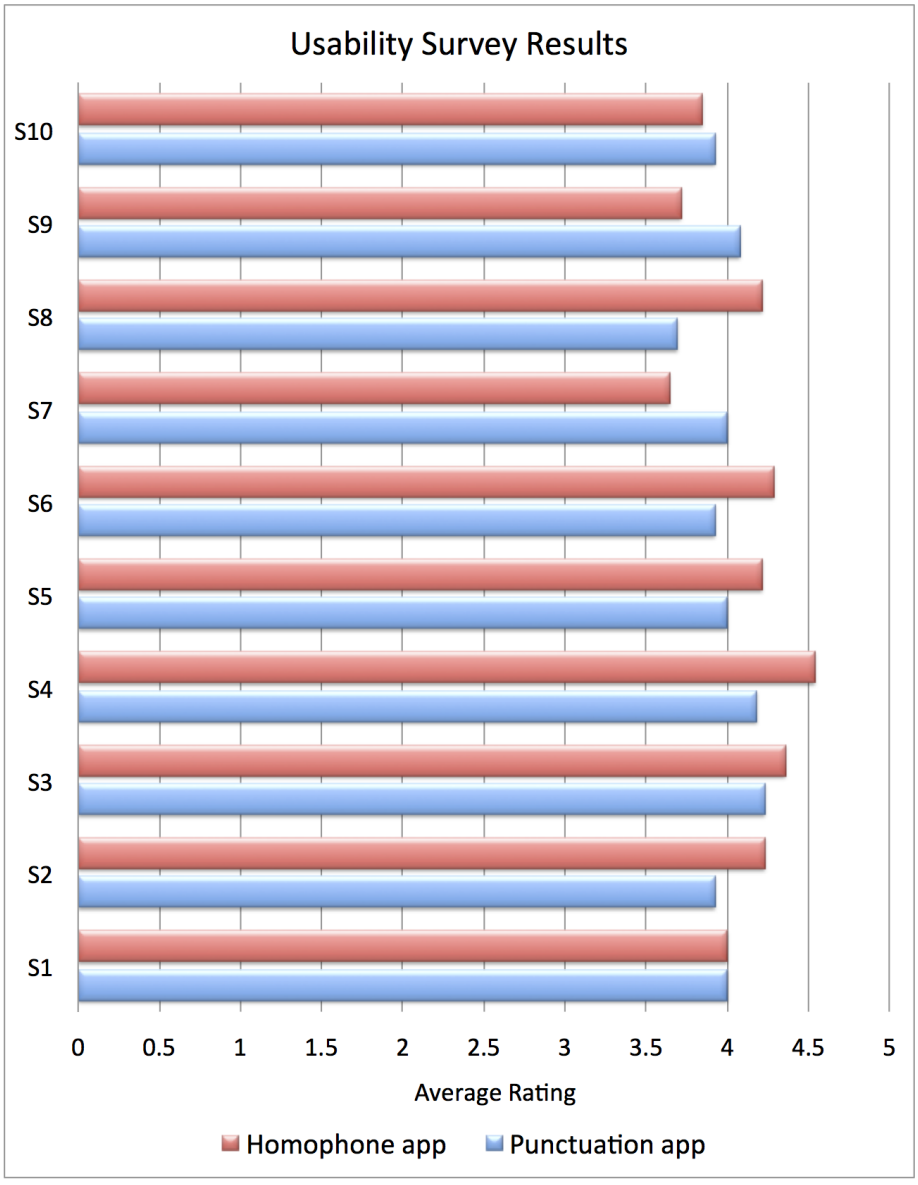

Figure 4.15: User experience survey results

The layered goals provided by the green and gold check marks were very effective at motivating user behavior as participants genuinely became emotionally involved in obtaining the correct answers. Exclamations like "Oh come on!" and "This is fun!" were recorded during the sessions, as well as laughter. Virtually every participant continued to work through the practice feature of each homophone submenu in the app until they had earned a gold check mark.

Though the homophone app and experiment were not designed with a social aspect in mind, by virtue of the fact that participants were sitting around the same large table, the participants' interactions with one another played a role. Several of the participants playfully competed against one another to complete the app by obtaining the gold check marks. Several participants also tilted their iPads so that they could see their neighbors screens and help each other out. Every participant discussed *some* aspect of the app while using the app with either the other participants, Gosse, or Browne.

One negative observation was initial frustration experienced by several users at what to do with the app from the beginning title screen. They felt confused by the amount of choice that they were provided, without any direction as to what to do. The confusion was momentary as we either explained what to do or the participants eventually experimented with the app by tapping one of the six buttons and entering a homophone submenu. Participants told us that this confusion could have been alleviated with audio providing direction.

Another negative observation was one user's frustration that the practice feature would force them to redo sentences that they had already got right during a previous attempt. They were stuck on the last of the five practice questions and were having to repeat all of the previous questions to keep trying to earn the gold check mark.

The results of the three homophone quizzes are shown in Figure 4.16, with separated average scores from the two sessions and the combined average. Overall in both sessions, the average quiz score slightly improved after both instructional methods.

The post-experiment questionnaire results regarding the preferred learning method and suggested future learning method(s) are found in Table 4.1 and Table 4.2. That the majority of participants suggested that in the future homophones be taught using both traditional methods and the app fits very well with observations and post-experiment discussions with participants.

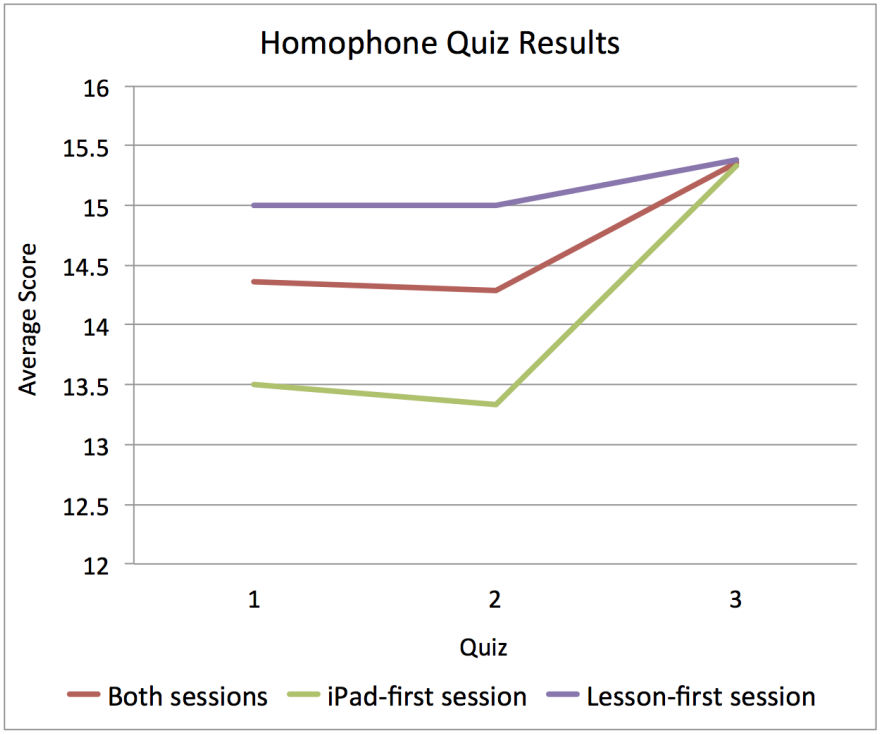

Figure 4.16: Homophone quiz results

Instruction method	Homophone sessions	Punctuation sessions
iPad	7	7
Lesson	7	6

Table 4.1: Preferred learning method

Instruction method	Homophone sessions	Punctuation sessions
iPad	0	0
Lesson	2	2
Both iPad & lesson	12	11

Table 4.2: Suggested future learning method

Though perhaps it may seem an obvious point, co-researcher Browne found it remarkable observing just how effective a human instructor can be at teaching literacy. Many participants had unique challenges to learning, whether it was their hearing ability, ability to stay focused, difficulty in grasping a concept, etc. As an instructor, Gosse was able to customize her approach to each individual participant's learning needs, from moment to moment as the situation evolved, in an emotionally sensitive manner. These abilities are very difficult for any piece of software to ever fully duplicate. The importance of these abilities during instruction may be even more important in a domain such as literacy that involves natural language and human-to-human communication by nature of the subject.

Discussions with the participants revealed that while many preferred to be taught the concepts via a traditional lesson, many preferred to practice and reinforce their understanding of the concepts with the iPad app. This view likely explains why the majority of participants suggested using both instructional methods in the future.

The discussion also revealed a strong consensus that after the users suffered an incorrect answer during the practice feature, the app should explain the correct answer to the user rather than displaying symbols that indicate an incorrect answer as the participants found this discouraging.

The results of this experiment prompted us to create the comma app, which as discussed in Section 4.3.3 follows a simple design similar to the practice feature of the homophone app. The comma app explains correct answers and does not force users to complete answers to already correct questions by instead cycling through the questions until they are all answered correctly.

The results for each individual participant are provided in Table 4.3, broken down into the session that started with the traditional method and the session that started with the iPad app. We note that of the six participants that rated their ability with the iPad as a '1', only one of these participants preferred the iPad app to the traditional lesson. The two participants who suggested that only the traditional lesson be used in the future also rated their ability with the iPad as a '1'. In contrast the three participants that rated their ability with the iPad as '4' or higher all preferred the iPad to the traditional lesson. Though more participants would be required to make stronger claims, this data provides some indication that a participant's ability to use the iPad may impact their instructional preference.

Regarding the relationship between quiz performance and instructional method preference, we observe that high performing participants P1, P3, P4,

Traditonal method first session								
P	**RA**	**iA**	**Q1**	**Q2**	**Q3**	**US**	**PM**	**SFM**
P1	4	1	17	17	17	4.1	Lesson	Both
P2	3	1	7	8	11	3.1	Lesson	Lesson
P3	5	1	16	17	17	3.3	Lesson	Lesson
P4	4	1	14	16	18	3.9	Lesson	Both
P5	4	1	16	15	14	4.7	iPad	Both
P6	4	3	17	18	17	4.6	Lesson	Both
P7	4	4	19	19	19	4.2	iPad	Both
P8	5	5	14	10	10	5.0	iPad	Both
iPad method first session								
P	**RA**	**iA**	**Q1**	**Q2**	**Q3**	**US**	**PM**	**SFM**
P9	2	1	4	1	5	1.7	Lesson	Both
P10	4	2	15	14	16	4.7	iPad	Both
P11	5	2	13	8	15	4.5	iPad	Both
P12	5	3	16	19	18	4.6	iPad	Both
P13	5	2	15	19	19	4.1	Lesson	Both
P14	5	5	18	19	19	4.7	iPad	Both

Table 4.3: Homophone experiment participant results - P = Participant, RA = Reading ability, iA = iPad ability, Q# = Quiz #, US = User experience score, PM = Preferred method, SFM = Suggested future method

P6 and P13 preferred the lesson, and high performing participants P7, P12 and P14 preferred the iPad app. This may indicate that quiz performance was not related to instructional method preference, however we also observe that lower performing participants P2 and P9 were also responsible for the lowest user experience score results (and both preferred the traditional lesson).

4.5.2 Punctuation Experiment

The punctuation app also received overall positive user experience feedback as shown in the survey results in Figure 4.15. In stark contrast to how the homophone app was discussed while it was being used, very little was spoken by participants while they were using the punctuation app. A couple of participants had difficulties that will be discussed, and they vocalized those

concerns while using the app, but the remaining participants largely remained silent until they had completed the app. This would be consistent with the participants experiencing a sensation of flow as the app intended. The fact that the app explained the correct answer if the user failed to pass a screen was well received, as was the clear direction as to what the app expected the user to do next.

Participants P1 and P6 (see Table 4.4) in particular had a very difficult time completing the app, for reasons that could have been alleviated had the design been different. In both cases, participants had rated their ability to use the iPad as "1" and were struggling to understand the concepts themselves. One of the participants had particular difficulty dragging the punctuation marks across the screen. The participant was lifting their finger from the screen while dragging the punctuation mark, and it would move back into its original position. Each time the user failed to pass a screen, they were given the exact same amount of time as before to pass the screen on the next try. Even when these participants knew the correct answer, they simply couldn't move the punctuation marks quickly enough. They would then become more frustrated when having to sit through the same explanation of the correct answer, and have to go through the process again.

Giving participants more time on "retries" and giving them different (but perhaps equally difficult) problems to work through could likely have prevented this frustration. Giving users more time or different problems could have been accomplished statically with different optional levels of difficulty at the start of the app, or dynamically as the user experienced the app. Also, instead of having the punctuation marks snap back to their original position after being placed in an incorrect position, they could have remained in the new but incorrect position, reducing this frustration as well.

Though less evident than the homophone app, again the relative dynamism of the instructor relative to the app was very apparent. Whether learning during the lesson or with the app, when participants failed to grasp something or get the right answer, they were understandably frustrated. In most cases by explaining the correct answer to the participant, the app was able to quell this frustration. But as noted above, sometimes this wasn't enough. Whereas the instructor was almost immediately aware of any frustration and able to deal with it appropriately, in a manner tailored to the participant.

During the punctuation sessions, the quiz results as presented in Figure 4.17 show an initial increase in score after the first instructional method (either traditional lesson or iPad) and a decrease after the second method. Perhaps some participants became bored after learning the same skills twice in a row.

As with the homophone app, the post-experiment questionnaire found

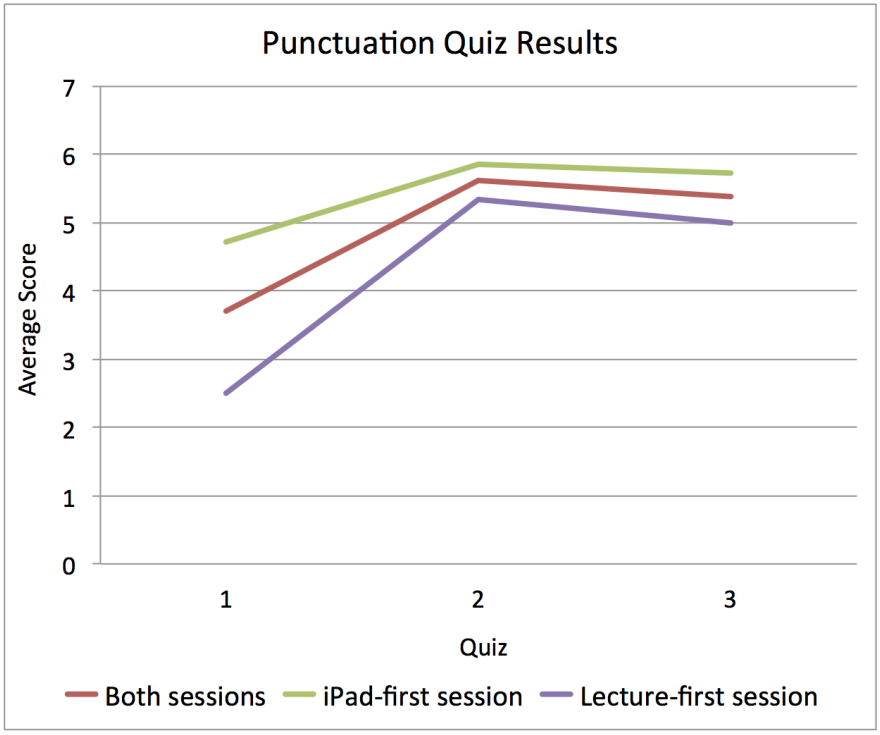

Figure 4.17: Punctuation quiz results

in Table 4.1 and Table 4.2 suggest that while about half of the participants preferred each method, there was near consensus that both methods should be used in the future.

The results for each individual participant are provided in Table 4.4, again broken down into the session that started with the traditional method and the session that started with the iPad app. Note that these participant numbers do not represent the same participants as the homophone app sessions presented in Table 4.3.

Two of the four participants who rated their ability with the iPad as a '1' preferred the traditional lesson to the iPad app. Five of the eight participants who rated their ability with the iPad as '4' or higher preferred the iPad app. Again this shows further studies with more participants would be required to make strong arguments about the relationship between the ability of the

Traditonal method first session								
P	**RA**	**iA**	**Q1**	**Q2**	**Q3**	**US**	**PM**	**SFM**
P1	3	1	1	2	1	3.0	iPad	Both
P2	5	4	6	7	7	4.0	iPad	Both
P3	4	4	0	8	5	4.8	iPad	Both
P4	3	5	2	4	5	4.0	iPad	Both
P5	5	5	6	10	10	4.1	Lesson	Both
P6	1	1	0	1	2	2.1	Lesson	Lesson
iPad method first session								
P	**RA**	**iA**	**Q1**	**Q2**	**Q3**	**US**	**PM**	**SFM**
P7	4	3	6	5	5	1.7	Lesson	Both
P8	5	5	2	2	2	4.7	Lesson	Both
P9	4	4	6	10	10	4.5	iPad	Both
P10	5	5	3	3	3	4.6	Lesson	Lesson
P11	4	1	3	4	3	4.1	iPad	Both
P12	5	1	9	10	10	4.7	Lesson	Both
P13	4	5	4	7	7	4.7	iPad	Both

Table 4.4: Punctuation experiment participant results - P = Participant, RA = Reading ability, iA = iPad ability, Q# = Quiz #, US = User experience score, PM = Preferred method, SFM = Suggested future method

participant to use the iPad and instructional method preference. The same is true of establishing relationships between quiz performance and instructional method performance, and other finer grain relationships of interest.

4.5.3 Comma App Experiment

The level of agreement with the first question of the comma-experiment questionnaire, "I liked using the app to do practice exercises during the lesson", was 4.4, indicating the app was well received. Of the 13 participants in the comma-app experiment, 5 reported that in the future they would prefer to do practice questions during lessons using the app, 6 said they would prefer either iPad or pen-and-paper practice questions, and 2 said they would prefer pen-and-paper practice questions.

Participants P1 and P6 (see Table 4.5) that said they would prefer pen

and paper were the same two participants that rated their iPad ability as "1" and struggled with the homophone app. Though not experiencing as much frustration with the homophone app, both participants were again frustrated with the comma app. Again difficulty in moving the object across the screen due to lifting the finger off the screen was an issue, as well as frustration due to having the same question repeated again and again after answering it incorrectly.

Other than these instances the app was well received, with the cycling of questions and explanation of the correct answers being particularly well received.

Traditonal method first session				
P	**RA**	**iA**	**SQ**	**FPM**
P1	3	1	3	Paper
P2	5	4	5	Either
P3	4	4	5	iPad
P4	3	5	4	Either
P5	5	5	4	Either
P6	1	1	3	Paper
iPad method first session				
P	**RA**	**iA**	**SQ**	**FPM**
P7	4	3	4	iPad
P8	5	5	5	iPad
P9	4	4	5	Either
P10	5	5	5	iPad
P11	4	1	4	Either
P12	5	1	5	Either
P13	4	5	5	iPad

Table 4.5: Comma experiment participant results - P = Participant, RA = Reading ability, iA = iPad ability, SQ = Survey likert question ('liked using the app'), FPM = Future preferred method

The results for each individual participant are provided in Table 4.5, again broken down into the session that started with the traditional method and the session that started with the iPad app. Note that these participant numbers are the same participants as the punctuation app sessions presented

in Table 4.4, as the brief experiment with the comma app was done as an addendum to the punctuation experiment sessions.

4.6 Conclusion

We conclude that integrating educational software incorporating game design elements into adult literacy education does increase learner engagement, on the basis of the participants' recommendation that the apps be used as part of future instruction in addition to traditional instruction. In addition, the participants' social activity and verbal expressions suggesting emotional involvement while using the homophone app, and the participants' relative silence while focused on using the punctuation app, are both evidence of contrasting forms of user engagement.

The gamification design approach of the homophone app facilitated an externally focused form of engagement. The layered goals of the app could be achieved at the participant's leisure as there was no time constraint. This allowed for participants to express their emotional state to one another, to compete with one another, and to help one another.

The serious game design approach of the punctuation app facilitated an internally focused form of engagement. The time limit kept participants focused on the challenge at hand, and the increasing difficulty of the challenges discouraged participants from becoming bored.

The external focus facilitated by the gamification design approach is consistent with the idea that gamification is about providing external rewards for the intrinsic motivations of users. The internally focused form of engagement facilitated by the serious game design approach is consistent with the idea that experiencing flow involves the loss of reflective self-consciousness as users become absorbed in the activity at hand. In either case participants were focused and engaged in the educational activity. We suggest both forms of engagement may be effective, depending on the desired approach and goals of the instruction.

The number of participants in this study does not allow us to generalize the more nuanced observations and results to the entire low-literacy adult population. For example, although it is encouraging that performance generally went up by the end of the homophone and punctuation sessions, the quiz results don't provide a strong enough signal to make broad claims about the educational effectiveness of one method or the other. For this, a further study with more participants or taking place over a larger amount of time would be necessary.

Our work and results provide re-confirmation of several user interface design concepts outlined previously in the literature (albeit in a new problem domain - adult literacy educational software):

- Use different levels of rewards (short, medium, and long term) to motivate behavior[HE11].
- Use time limits and gradually increasing challenge difficulty (both in time allotted and conceptual difficulty) to encourage the user to experience flow during challenge(s)[KBBK11].
- Immediately present a new challenge after the user has completed their attempts at the current challenge to encourage the user to experience flow[JW03, SRV13].
- Don't present a user with the same challenge if they have already failed the challenge twice[KMP99].

This work was motivated by the high economic and personal costs of low adult literacy and the potential for tablet educational software based on gamification and serious game design approaches to offer novel solutions that may increase engagement, improve cost effectiveness, be easier to use, and / or have the ability to be used remotely at the user's home.

The majority of participants were able to use the software without experiencing too much frustration, and it appears that design improvements could have prevented the frustration that was experienced. This indicates that ease of use is indeed a positive factor for using tablets for adult literacy education.

While a majority of participants suggested the apps be used when the concepts are taught in the future, this was in addition to a traditional lesson. If a human instructor is still required, this limits the potential to improve cost effectiveness, or the ability of the user to benefit from being able to learn at home.

Regarding cost effectiveness, an instructor naturally has a limited amount of time that can be allotted to some mix of lessons, one-on-one tutoring, marking practice questions, etc. The positive reception of the homophone app's practice feature and the comma app suggests that instructors could automate at least some of the time spent marking practice questions using tablet apps. Using educational apps incorporating gamification for drill and practice would correspond with the historical role of video games in the classroom[Squ03].

Landers and Callan[LC11] use a hammer and nail analogy to suggest educators pick a game as an instructional method only *after* they have specified

what concepts are to be instructed. If adult literacy instructors could easily pick through suites of effectively designed apps made to drill and practice specific literacy concepts, they could use them to replace this portion of their time, increasing the cost effectiveness of literacy instruction.

The fact that participants suggested both a human instructor and an iPad app be used in future instruction of the concepts doesn't mean the concepts couldn't be learned from home via the apps if necessary. But looking towards future work in this area, a holy grail for literacy educational software would be completely automated instruction of all concepts done at such a high quality that learners *prefer* the apps to a human. This may very well be impossible, but we suggest a design approach that may aid in moving towards this goal.

In our experiments, it was the dynamism of the instructor that appeared to make them so effective at teaching the concepts themselves. In-the-moment adjustments based on participant reactions, learning tailored to individual needs, and emotional sensitivity all played a clear role in making the instructor effective in this way.

Our apps came closest to this sort of dynamism when they themselves were dynamic. The stair-climbing flow-based design of the punctuation app allowed for advanced individuals to proceed quickly through the app, with jumps in challenge difficulty that were accommodating enough for most participants. There was also a strong positive reception of other "dynamic" features like explanations of correct answers after suffering an incorrect answer.

We suggest that a high degree of dynamic behavior in future adult-literacy educational apps will have a positive effect on their reception and ability to replace instructors in roles beyond drill and practice. In particular, we put forward a concept of *dynamic flow inducement* for exploration.

The concept of flow offers promise in educational software beyond drill and practice because by nature it involves the stretching of skills to meet new challenges, which can then be stretched further and further, to build up a more sophisticated understanding over time. The problem with using flow in this "stair-climbing to understanding" strategy is that users all start off from different levels of understanding and abilities, and will proceed in their growth of understanding at different rates. An app could account for these issues and sustain an experience of flow by dynamically adjusting the difficulty of the challenges, the way the challenges are presented, the time requirements, etc. We propose that this concept be explored with more extensive experiments designed to test the effectiveness of strategies for maintaining an experience of flow while dynamically adjusting the challenges presented to the user.

4.7 Bibliography

[A+10] David Autor et al. The polarization of job opportunities in the us labor market: Implications for employment and earnings. *Center for American Progress and The Hamilton Project*, 2010.

[Bar12] Frank Barry. How online fundraising, gamification and social media helped raise over $2 million dollars in one day, 2012. http://www.npengage.com/onlinefundraising/howonlinefundraisinggamificationandsocialmediahelpedraiseover2milliondollarsoneday/ (accessed April 5, 2013).

[BTNP05] Lynn Barr-Telford, François Nault, and Jean Pignal. Building on our competencies: Canadian results of the international adult literacy and skills survey. *Statistics Canada. Available at: www. statcan. ca/bsolc/english/bsolc*, 2005.

[CG11] Alma L Culén and Andrea Gasparini. ipad: a new classroom technology? a report from two pilot studies. *INFuture Proceedings*, pages 199–208, 2011.

[Che12] Brian X. Chen. Hmh fuse pilot program, 2012. http://www.hmheducation.com/fuse/pilot-1.php.

[Csi97] Mihaly Csikszentmihalyi. *Finding flow: The psychology of engagement with everyday life*. Basic Books, 1997.

[DDKN11] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, MindTrek '11, pages 9–15, New York, NY, USA, 2011. ACM.

[Det11] S. Deterding. Situated motivational affordances of game elements: A conceptual model., 2011. Presented at Gamification: Using Game Design Elements in Non-Gaming Contexts, a workshop at CHI 2011. Retrieved from http://gamification-research.org/wp-content/uploads/2011/04/09-Deterding.pdf.

[DM12] Alec Dorling and Fergal McCaffery. The gamification of spice. In *Software Process Improvement and Capability Determination*, pages 295–301. Springer, 2012.

[DMT05] Richard Desjardins, TS Murray, and AC Tuijnman. *Learning a living: First results of the Adult Literacy and Life Skills survey*. OECD, 2005.

[Gar12] Gartner. Gartner says by 2014, 80 percent of current gamified applications will fail to meet business objectives primarily due to poor design, 2012. http://www.gartner.com/newsroom/id/2251015 (accessed April 5, 2013).

[Gar13] Gartner. Gartner says worldwide pc, tablet and mobile phone combined shipments to reach 2.4 billion units in 2013, 2013. http://www.gartner.com/newsroom/id/2408515 (accessed April 5, 2013).

[Gen90] James W Gentry. What is experiential learning. *Guide to business gaming and experiential learning*, pages 9–20, 1990.

[HE11] Juho Hamari and Veikko Eranti. Framework for designing and evaluating game achievements. *Proc. DiGRA 2011: Think Design Play*, 115:122–134, 2011.

[Hig08] Reach Higher. America: Overcoming the crisis in the us workforce, 2008.

[HY12] S. Henderson and J. Yeow. ipad in education: A case study of ipad adoption and use in a primary school. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 78–87, 2012.

[JW03] Daniel Johnson and Janet Wiles. Effective affective user interface design in games. *Ergonomics*, 46(13-14):1332–1345, 2003.

[KBBK11] Johannes Keller, Herbert Bless, Frederik Blomann, and Dieter Kleinböhl. 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, 2011.

[KGB05] M Kutner, E Greenberg, and J Baer. National assessment of adult literacy (naal): A first look at the literacy of americas adults in the 21st century (report no. nces 2006–470). *Washington, DC: National Center for Education Statistics, US Department of Education*, 2005.

[KMP99] Jonathan Klein, Youngme Moon, and Rosalind W Picard. This computer responds to user frustration. In *CHI'99 extended abstracts on Human factors in computing systems*, pages 242–243. ACM, 1999.

[LC11] Richard N. Landers and Rachel C. Callan. Casual social games as serious games: The psychology of gamification in undergraduate education and employee training. In Minhua Ma, Andreas Oikonomou, and Lakhmi C. Jain, editors, *Serious Games and Edutainment Applications*, pages 399–423. Springer London, 2011. 10.1007/9781447121619_20.

[Lit01] BC Literacy. *Who Wants to Learn?: Patterns of Participation in Canadian Literacy and Upgrading Programs: Summary of Results from the National Study Conducted for ABC Canada in Partnership with Literacy BC.* ABC Canada, 2001.

[Lon01] Ellen Long. *Patterns of Participation in Canadian Literacy and Upgrading Programs: Results of a National Follow-Up Study.* ERIC, 2001.

[MRSD12] Rajesh Kannan Megalingam, Ananthakrishnan P Rajendran, Abhiram T Solamon, and Deepak Dileep. Edupada tablet based educational system for improving adult literacy in rural india. In *Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on*, pages 1–5. IEEE, 2012.

[NC09] Jeanne Nakamura and Mihaly Csikszentmihalyi. Flow theory and research. *Handbook of positive psychology*, pages 195–206, 2009.

[REZ03] Stephen P. Rushton, Janice Eitelgeorge, and Ruby Zickafoose. Connecting brian cambourne's conditions of learning theory to brain/mind principles: Implications for early childhood educators. *Early Childhood Education Journal*, 31:11–21, 2003. 10.1023/A:1025128600850.

[Squ03] Kurt D. Squire. Video games in education. *Int. J. Intell. Games & Simulation*, 2(1):49–62, 2003.

[SRV13] Jorge SimíEs, Rebeca DíAz Redondo, and Ana FernáNdez Vilas. A social gamification framework for a k-6 learning platform. *Comput. Hum. Behav.*, 29(2):345–353, March 2013.

[Tou11] Kayne Toukonn. The dynamic electronic textbook: Enhancing the student's learning experience. *Kent State University*, 2011. MFA Thesis.

[VTRB07] J Vernon, A Trujillo, S Rosenbaum, and DeBuono B. Low health literacy: implications for national health policy. *University of Connecticut. National Bureau of Economic Research, Storrs, CT*, 2007.

[WCD11] K. Wattanatchariya, S. Chuchuaikam, and N. Dejdumrong. An educational game for learning wind and gravity theory on ios: Drop donuts. In *Computer Science and Software Engineering (JCSSE), 2011 Eighth International Joint Conference on*, pages 387 –392, may 2011.

[Yan11] Feng Yan. A sunny day: Ann and rons world an ipad application for children with autism. In Minhua Ma, Manuel Fradinho Oliveira, and Joo Madeiras Pereira, editors, *Serious Games Development and Applications*, volume 6944 of *Lecture Notes in Computer Science*, pages 129–138. Springer Berlin / Heidelberg, 2011.

[YJM+09] H Shonna Yin, Matthew Johnson, Alan L Mendelsohn, Mary Ann Abrams, Lee M Sanders, and Benard P Dreyer. The health literacy of parents in the united states: a nationally representative study. *Pediatrics*, 124(Supplement 3):S289–S298, 2009.

[Zam12] Zamzee. New research shows zamzee increases physical activity by almost 60%, 2012. http://blog.zamzee.com/2012/09/26/newresearchshowszamzeeincreasesphysicalactivitybyalmost60/ (accessed April 5, 2013).

Chapter 5

Reading Comprehension App Study

The following work is not yet published as of August 2015. We plan to submit this work to a peer-reviewed journal or conference.

5.1 Abstract

Motivated by the substantial health, economic and employment issues associated with low literacy skills, we developed a tablet application to help adults improve their reading comprehension skills. In this paper, we overview its design and present the results of an evaluation performed with post-secondary students. Though reading comprehension is a complex and multifaceted skill, studies have shown that teaching students metacognitive techniques can help them improve their reading comprehension skills. We created an iPad application which attempts to teach users the question generation strategy, while using gamification design approaches, in particular dynamic difficulty adjustment, to induce a high level of user engagement. We built another tablet application as a control that only allowed users to practice their reading comprehension skills; this application lacked most of the gamification elements and 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 also found that while some gamification design elements were effective, participants desired improvements to the implementation of dynamic difficult adjustment. We

conclude that tablet software can be used to teach the question generation strategy, and propose directions for future work.

5.2 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 [Hig08, Lit01, VTRB07, YJM+09, BTNP05, KGB05]. 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%[Lon01], and under 10% of Canadians who could benefit from a literacy program enroll[Lit01], with those who do not enroll citing employment, financial, childcare and transportation constraints.

Tablet computers have exploded in popularity, reaching 116 million in sales worldwide in 2012 and are projected to grow to 468 million by 2017[Gar15]. 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[App88].

This opportunity led the authors to conduct an exploratory study into the effectiveness of tablet software incorporating gamification and serious game design approaches[BAG14]. 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 Cen-

tre instructions, and incorporated Brian Cambourne's and Kayne Toukonen's thoughts on learning[REZ03, Tou11], 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.

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[DC80], vocabulary knowledge[COL04], background knowledge[PHG79], phonology[BA+90], interest level[SO96], inference-making ability[CO98], text-anomaly resolution ability[YOP89], and cultural background[Joh81]. 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[SS94]. 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[LFRB95]. 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[SCC10, McN12, JD04]. One way that proficient readers are different from struggling readers is in their application of metacognitive reading strategies[BB84]. 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[BL91] or summarize[BS84] a passage of text. Proficient readers will employ these strategies before, after, and during the reading of a passage of text[PWT91]. 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[SCC10, WJ82, BB85]. Some experimental results have shown that struggling readers can improve their reading comprehension in a single learning session[GB86, BB85]. While reading comprehension strategies have previously been taught using software successfully, for example iSTART[MLB04], we are unaware of any study documenting using tablet software to teach a reading comprehension strategy.

The *question generation* reading comprehension strategy involves having the learner generate and answer questions in the process of reading the text[Coh83, Ros97]. 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[Hun05], 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"[Det11]. Dynamic difficulty adjustment fits this definition and can be used to make an experience engaging to a wide spectrum of different users[Mis14].

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?
- 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.

In Section 5.3, we present the design of the iPad applications. In Section 5.4, we discuss the design of the usability experiment that took place. In Section 5.5, we analyze the results of this study, and in Section 5.6, we discuss our conclusions and provide directions for future work.

5.3 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.

5.3.1 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 5.1. 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.

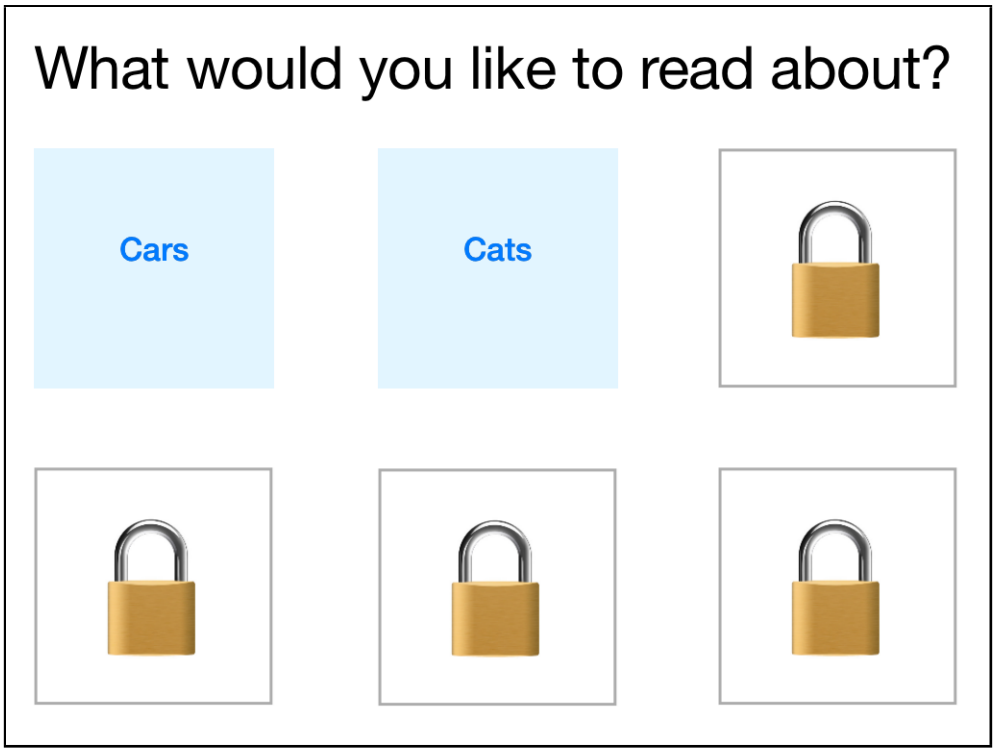

Figure 5.1: Topic selection screen

After the user selects a topic, they are presented with the text screen shown in Figure 5.2. 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.

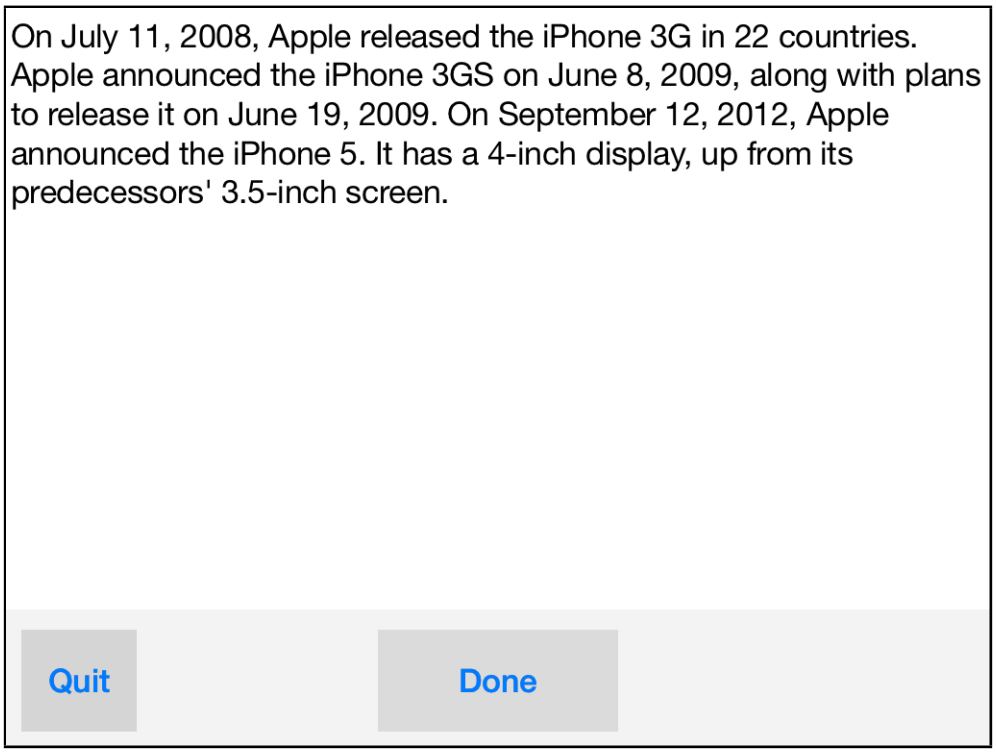

Figure 5.2: Text screen

After the user clicks "Done" on the text screen they are presented with the question screen found in Figures 5.3 and 5.4. 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 5.3. 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 5.4. 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).

Once the user has completed the set of questions associated with the

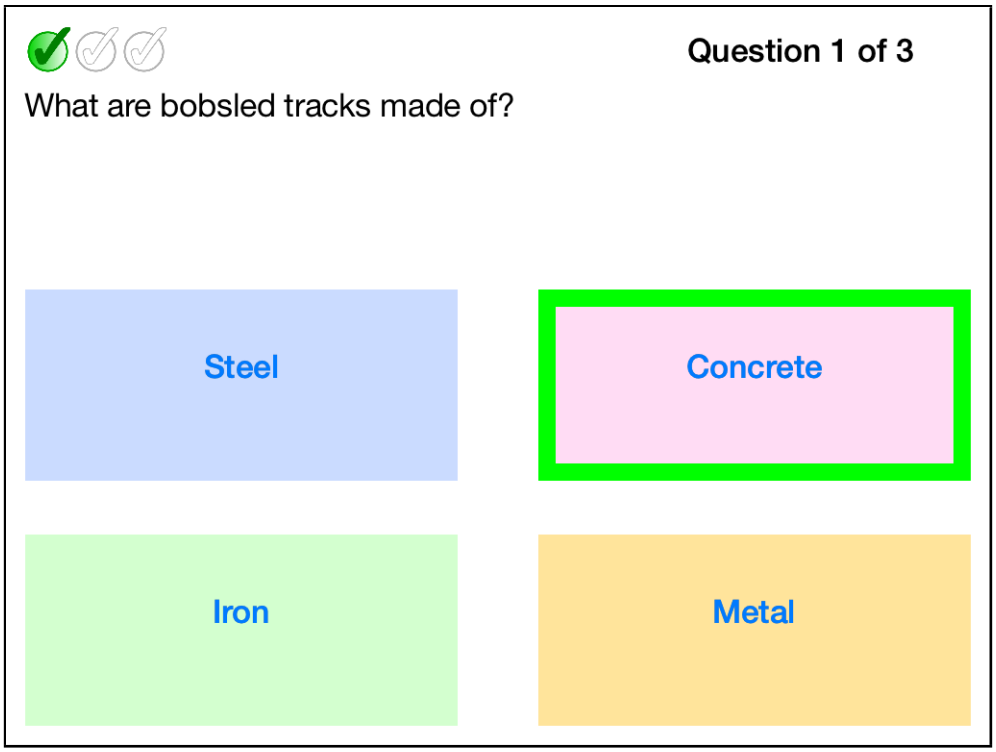

Figure 5.3: Correct answer

passage, they are forwarded to the reward screen shown in Figure 5.5. 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.

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

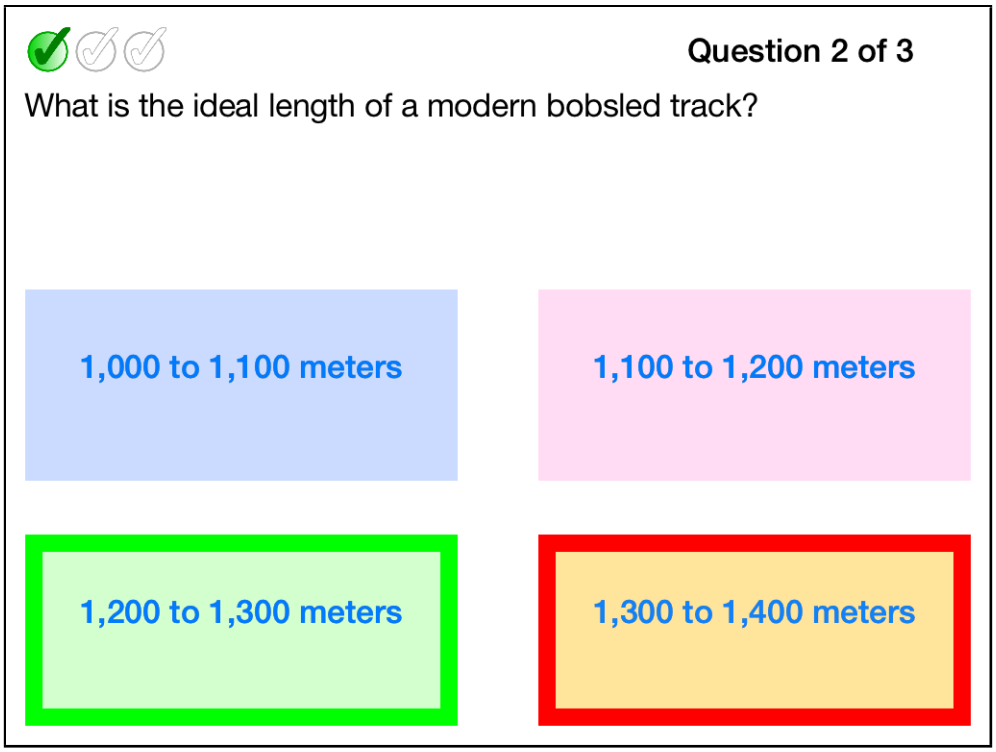

Figure 5.4: Incorrect answer

screens shown in Figures 5.6, 5.7 and 5.8.

The motivation screen shown in Figure 5.6 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.

After clicking "Next" on the motivation screen, the user is presented with an example screen as in Figure 5.7, 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.

After the user clicks "Next" in the example screen, the user is presented with the tactics screen shown in Figure 5.8. The tactics screen presents a tactic to help the user apply the question generation strategy, i.e. a method

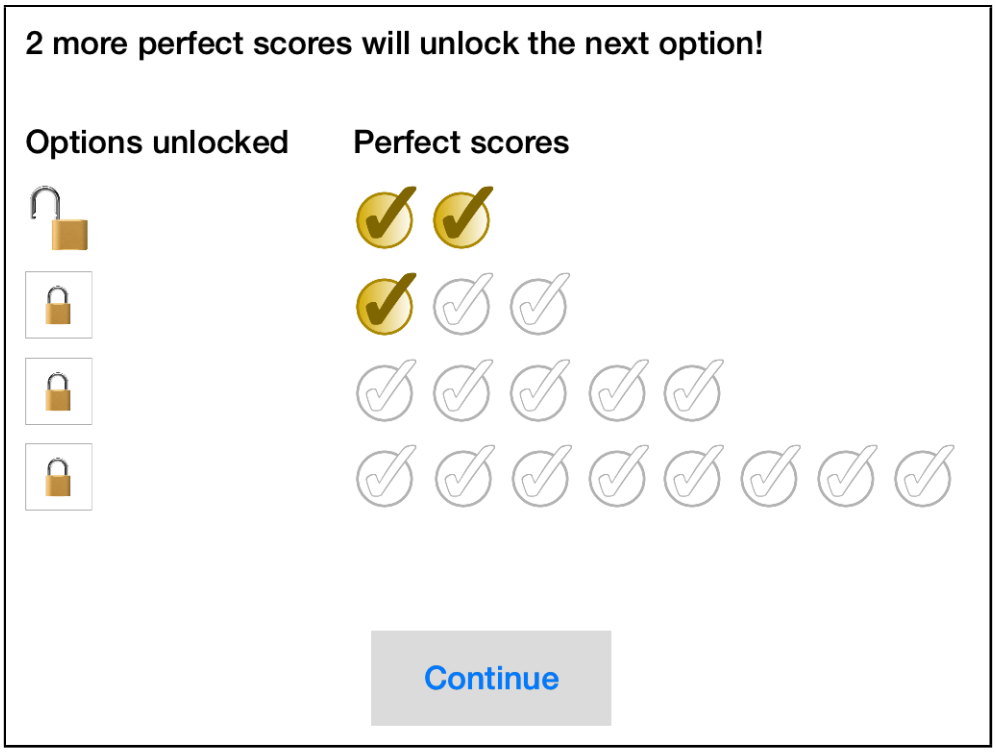

Figure 5.5: Reward screen

of carrying out the question generation strategy. The tactics screen randomly presents one of four possible tactics for carrying out the question generation strategy.

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.

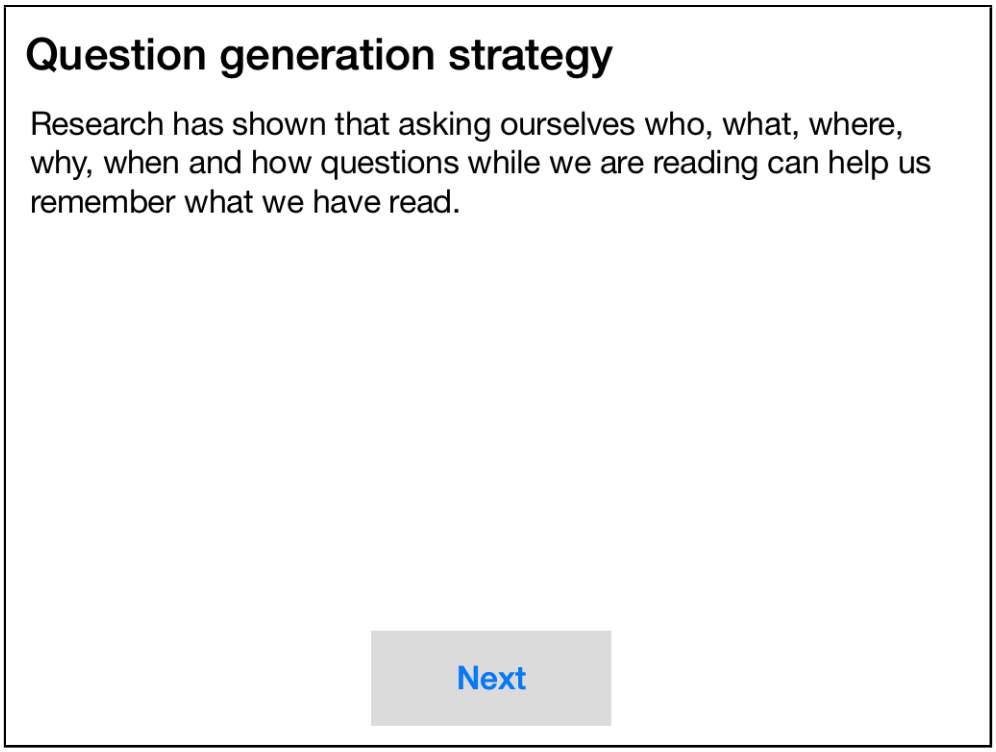

Figure 5.6: Motivation screen

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[KAOC81]. The levels and associated question count, word count range, and Flesch-Kincaid score range are shown in Table 5.1.

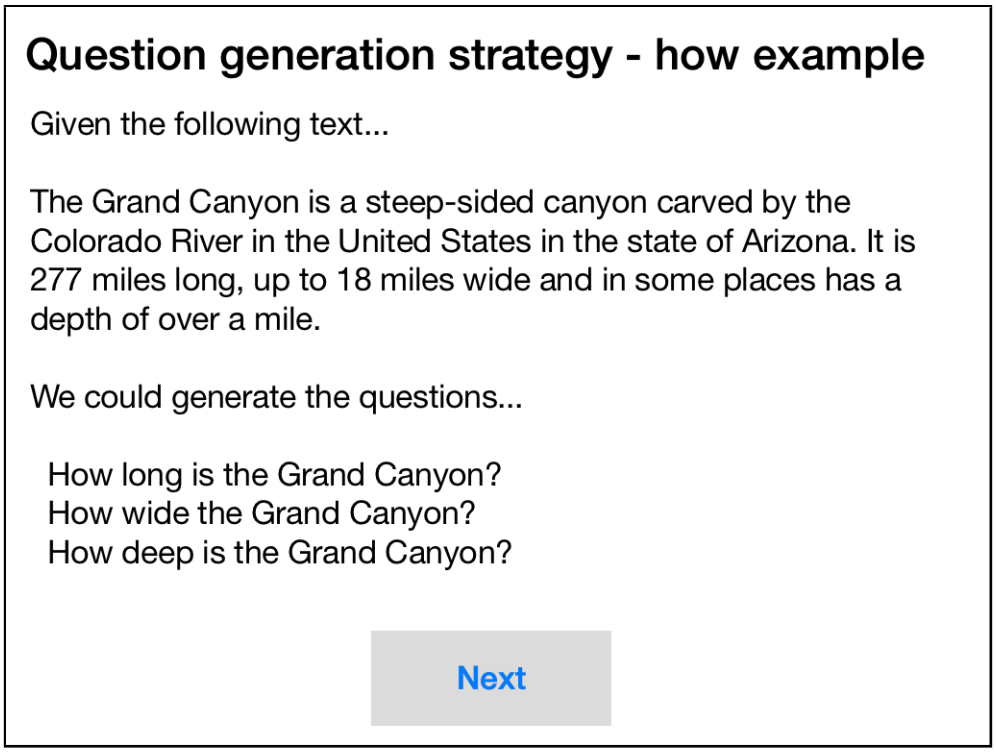

Figure 5.7: Example screen

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 5.1: Passage levels

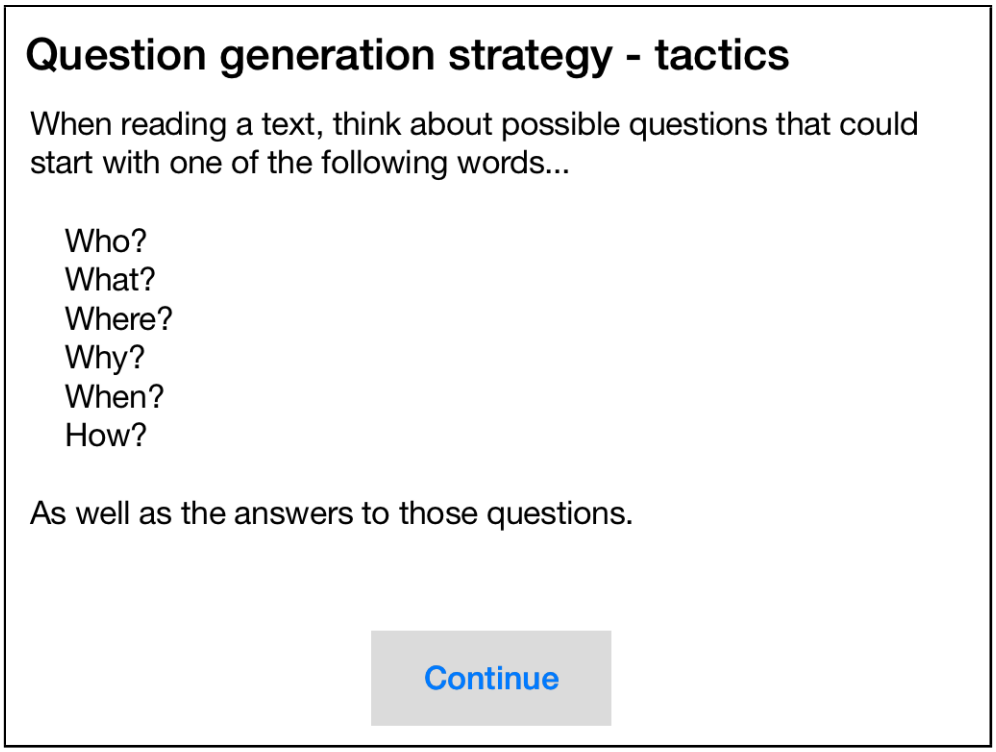

Figure 5.8: Tactic screen

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 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, chem-

istry) 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 [Coh83, Ros97]. Metacognitive strategies include three components: declarative ("knowing what"), procedural ("knowing how"), and conditional ("knowing why") [PLW83]. Teaching a metacognitive strategy can be done by providing explicit answers for these what, how and why questions[Car98]. 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[NC09] describe flow as a subjective experience that seamlessly unfolds from moment to moment. Csíkszentmihályi[Csi97] models flow as a balance between perceived opportunities and skills, with the current model of flow shown in Figure 5.9 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.

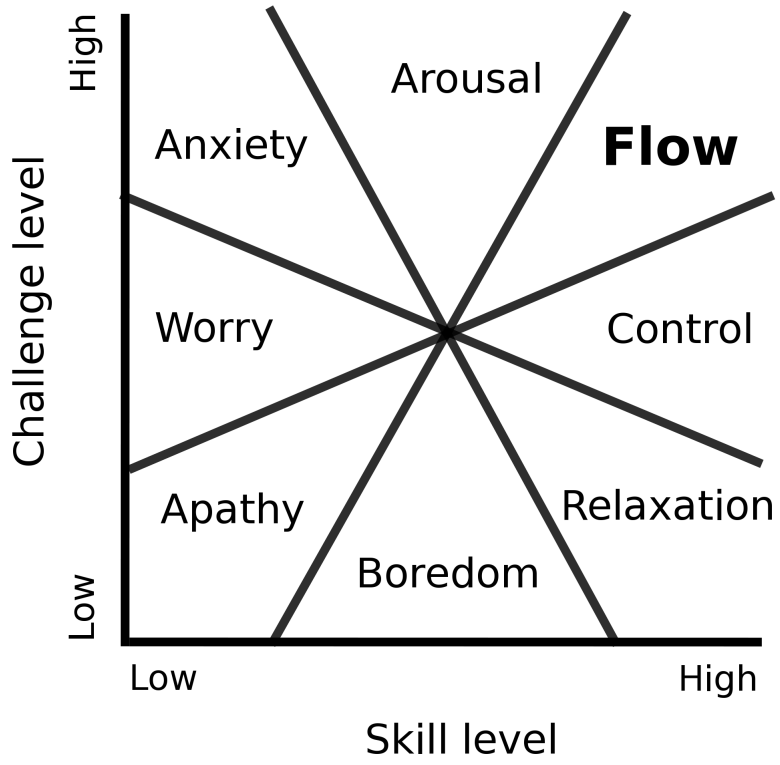

Figure 5.9: Csíkszentmihályi model of flow

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[DM12].

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[Kol85]. 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.

5.3.2 Control Application

The control application works by cycling only between the text screen in Figure 5.2, and the question screen in Figure 5.3 and Figure 5.4. 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.

5.3.3 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.

5.4 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 Research Ethics Board.

5.4.1 Experiment session protocol

The following protocol was followed with each experiment participant. The protocol refers to the pre-experiment questionnaire in Section 2.4.4, the usability survey in Section 4.4.5, and the post-experiment questionnaire in Section 2.4.6.

The sessions took place in a classroom and a meeting room at McMaster University and were conducted by Kevin 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[GB86].

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.

5.4.2 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.

5.4.3 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 4.4.5.

5.4.4 Pre-experiment questionnaire

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".

5.4.5 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.
- **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 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[BAG14]. Statements S11-S17 are modeled closely after those used in a survey to measure engagement in a prior study by Webster and Ho[WH97]. Statements S18-S23 are modeled closely after those used in a survey to measure flow in a prior study Choi and Kim[CK04]. As a result of combining multiple different survey instruments, some survey questions are very similar (e.g. S15 and S19).

5.4.6 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)

5.5 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 5.2. 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 5.2: 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 5.10, 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 $\alpha = 0.05$. 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:

Source	**SS**	**df**	**MS**	**F**
Between Subjects	453.24	47		
Factor$_{BS}$	19.26	1	19.26	2.04
Error	433.98	46	9.43	
Within Subjects	261.5	48		
Factor$_{WS}$	15.84	1	15.84	3.25
Factor$_{BSxWS}$	21.09	1	21.09	4.32
Error	224.57	46	4.88	
Total	714.74	95		

Table 5.3: ANOVA Summary Table

We present the results of the ANOVA in summary Table 5.3, where SS is the sum of squares, df is the degrees of freedom, MS is the mean square, and F is the test statistic.

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).

The result of the usability survey are presented in Figure 5.11. 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 4.4.5 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

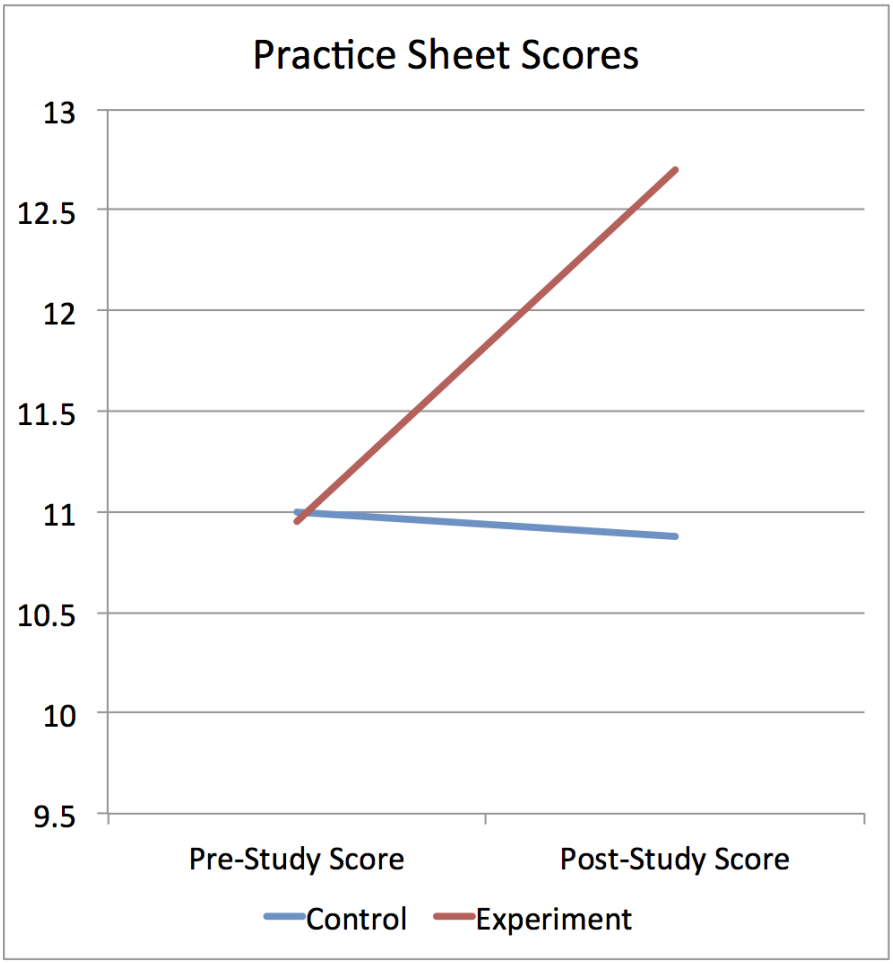

Figure 5.10: Practice sheet scores

case of S1-S10), and multiply the result by 100 to give a rough score. We present these results in Figure 5.12, where again we note that the experiment application has received higher results, but not significantly higher results as was thought possible.

In Table 5.4 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.

Figure 5.11: Usability survey results

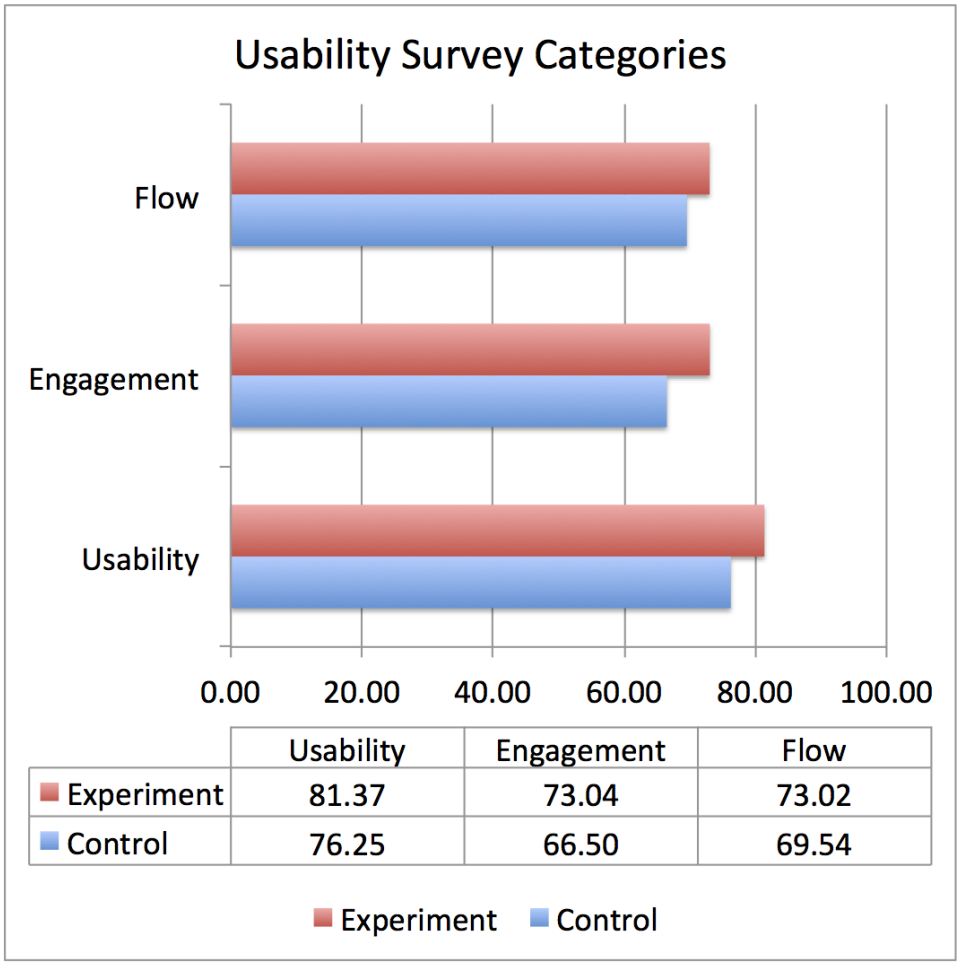

Figure 5.12: Usability survey categories

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

Table 5.4: Preferred learning method

In Table 5.5 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 5.5: 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 which emphasized the importance of hiding the existence of dynamic difficulty adjustment from the player for the technique to work as intended[RA03]. 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 5.6 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 5.6: Total passages read

5.6 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 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[HE11] and gradually increasing difficultly levels[KBBK11]) 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.

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.

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.

5.7 Bibliography

[App88] Apple. Apple vision, 1988. https://www.youtube.com/watch?v=oknObWgOrV4 (accessed May 4, 2015).

[BA+90] Dorothy VM Bishop, Catherine Adams, et al. 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, 1990.

[BAG14] Kevin Browne, Christopher Anand, and Elizabeth Gosse. Gamification and serious game approaches for adult literacy tablet software. *Entertainment Computing*, 5(3):135–146, 2014.

[BB84] Linda Baker and Ann L Brown. Metacognitive skills and reading. *Handbook of reading research*, 1(353):V394, 1984.

[BB85] Carl Bereiter and Marlene Bird. Use of thinking aloud in identification and teaching of reading comprehension strategies. *Cognition and instruction*, 2(2):131–156, 1985.

[BL91] Nanci Bell and Phyllis Lindamood. *Visualizing and verbalizing: For language comprehension and thinking*. Academy of Reading Publications Paso Robles, CA, 1991.

[BS84] Thomas W Bean and Fern L Steenwyk. The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. *Journal of Literacy Research*, 16(4):297–306, 1984.

[BTNP05] Lynn Barr-Telford, François Nault, and Jean Pignal. Building on our competencies: Canadian results of the international adult literacy and skills survey. *Statistics Canada. Available at: www. statcan. ca/bsolc/english/bsolc*, 2005.

[Car98] Patricia L Carrell. Can reading strategies be successfully taught? *Australian Review of Applied Linguistics*, 21:1–20, 1998.

[CK04] Dongseong Choi and Jinwoo Kim. 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, 2004.

[CO98] Kate Cain and Jane Oakhill. Comprehension skill and inference-making ability: Issues of causality. *Reading and spelling: Development and disorders*, pages 329–342, 1998.

[Coh83] Ruth Cohen. Self-generated questions as an aid to reading comprehension. *The Reading Teacher*, pages 770–775, 1983.

[COL04] Kate Cain, Jane Oakhill, and Kate Lemmon. 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, 2004.

[Csi97] Mihaly Csikszentmihalyi. *Finding flow: The psychology of engagement with everyday life*. Basic Books, 1997.

[DC80] Meredyth Daneman and Patricia A Carpenter. Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, 19(4):450–466, 1980.

[Det11] S. Deterding. Situated motivational affordances of game elements: A conceptual model., 2011. Presented at Gamification: Using Game Design Elements in Non-Gaming Contexts, a workshop at CHI 2011. Retrieved from http://gamification-research.org/wp-content/uploads/2011/04/09-Deterding.pdf.

[DM12] Alec Dorling and Fergal McCaffery. The gamification of spice. In *Software Process Improvement and Capability Determination*, pages 295–301. Springer, 2012.

[Gar15] Gartner. Gartner says tablet sales continue to be slow in 2015, 2015. http://www.gartner.com/newsroom/id/2954317 (accessed May 4, 2015).

[GB86] Linda B Gambrell and Ruby J Bales. Mental imagery and the comprehension-monitoring performance of fourth-and fifth-grade poor readers. *Reading Research Quarterly*, pages 454–464, 1986.

[HE11] Juho Hamari and Veikko Eranti. Framework for designing and evaluating game achievements. *Proc. DiGRA 2011: Think Design Play*, 115:122–134, 2011.

[Hig08] Reach Higher. America: Overcoming the crisis in the us workforce, 2008.

[Hun05] Robin Hunicke. The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pages 429–433. ACM, 2005.

[JD04] Tamara L Jetton and Janice A Dole. *Adolescent literacy research and practice*. Guilford Publications, 2004.

[Joh81] Patricia Johnson. Effects on reading comprehension of language complexity and cultural background of a text. *TESOL quarterly*, pages 169–181, 1981.

[KAOC81] J Peter Kincaid, James A Aagard, John W O'Hara, and Larry K Cottrell. Computer readability editing system. *Professional Communication, IEEE Transactions on*, (1):38–42, 1981.

[KBBK11] Johannes Keller, Herbert Bless, Frederik Blomann, and Dieter Kleinböhl. 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, 2011.

[KGB05] M Kutner, E Greenberg, and J Baer. National assessment of adult literacy (naal): A first look at the literacy of americas adults in the 21st century (report no. nces 2006–470). *Washington, DC: National Center for Education Statistics, US Department of Education*, 2005.

[Kol85] David Kolb. Learning styles inventory. *The Power of the 2 2 Matrix*, page 267, 1985.

[LFRB95] Diane Lapp, James Flood, and Wendy Ranck-Buhr. Using multiple text formats to explore scientific phenomena in middle school classrooms. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, 11(2):173–186, 1995.

[Lit01] BC Literacy. *Who Wants to Learn?: Patterns of Participation in Canadian Literacy and Upgrading Programs: Summary of Results from the National Study Conducted for ABC Canada in Partnership with Literacy BC.* ABC Canada, 2001.

[Lon01] Ellen Long. *Patterns of Participation in Canadian Literacy and Upgrading Programs: Results of a National Follow-Up Study.* ERIC, 2001.

[McN12] Danielle S McNamara. *Reading comprehension strategies: Theories, interventions, and technologies*. Psychology Press, 2012.

[Mis14] Olana Missura. Dynamic difficulty adjustment. 2014.

[MLB04] Danielle S McNamara, Irwin B Levinstein, and Chutima Boonthum. istart: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, 36(2):222–233, 2004.

[NC09] Jeanne Nakamura and Mihaly Csikszentmihalyi. Flow theory and research. *Handbook of positive psychology*, pages 195–206, 2009.

[PHG79] P David Pearson, Jane Hansen, and Christine Gordon. The effect of background knowledge on young children's comprehension of explicit and implicit information. *Journal of Literacy Research*, 11(3):201–209, 1979.

[PLW83] Scott G Paris, Marjorie Y Lipson, and Karen K Wixson. Becoming a strategic reader. *Contemporary educational psychology*, 8(3):293–316, 1983.

[PWT91] Scott G Paris, Barbara Wasik, and Julianne C Turner. The development of strategic readers. 1991.

[RA03] Andrew Rollings and Ernest Adams. *Andrew Rollings and Ernest Adams on game design*. New Riders, 2003.

[REZ03] Stephen P. Rushton, Janice Eitelgeorge, and Ruby Zickafoose. Connecting brian cambourne's conditions of learning theory to brain/mind principles: Implications for early childhood educators. *Early Childhood Education Journal*, 31:11–21, 2003. 10.1023/A:1025128600850.

[Ros97] Barak Rosenshine. The case for explicit, teacher-led, cognitive strategy instruction. *MF Graves (Chair), What sort of comprehension strategy instruction should schools provide*, 1997.

[SCC10] Hui-Fang Shang and I-Ju Chang-Chien. The effect of self-questioning strategy on efl learners reading comprehension development. *The International Journal of Learning*, 17(2):41–54, 2010.

[SO96] Ivor Sousa and Jane Oakhill. Do levels of interest have an effect on children's comprehension monitoring performance? *British Journal of Educational Psychology*, 66(4):471–482, 1996.

[SS94] Keith E Stanovich and Linda S Siegel. 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, 1994.

[Tou11] Kayne Toukonn. The dynamic electronic textbook: Enhancing the student's learning experience. *Kent State University*, 2011. MFA Thesis.

[VTRB07] J Vernon, A Trujillo, S Rosenbaum, and DeBuono B. Low health literacy: implications for national health policy. *University of Connecticut. National Bureau of Economic Research, Storrs, CT*, 2007.

[WH97] Jane Webster and Hayes Ho. Audience engagement in multimedia presentations. *ACM SIGMIS Database*, 28(2):63–77, 1997.

[WJ82] Bernice YL Wong and Wayne Jones. Increasing metacomprehension in learning disabled and normally achieving students through self-questioning training. *Learning Disability Quarterly*, 5(3):228–240, 1982.

[YJM+09] H Shonna Yin, Matthew Johnson, Alan L Mendelsohn, Mary Ann Abrams, Lee M Sanders, and Benard P Dreyer. The health literacy of parents in the united states: a nationally representative study. *Pediatrics*, 124(Supplement 3):S289–S298, 2009.

[YOP89] Nicola Yuill, Jane Oakhill, and Alan Parkin. Working memory, comprehension ability and the resolution of text anomaly. *British journal of psychology*, 80(3):351–361, 1989.

Chapter 6

Conclusion

In Section 6.1, we reexamine our hypothesis in light of the evidence; in Section 6.2, we provide a global analysis of findings across the studies conducted; and, in Section 6.3, present avenues for future work.

6.1 Hypothesis Confirmation

Our results provide *statistically significant confirmation* of the hypothesis that gamification design approaches can be used to create engaging and educationally effective mobile educational software.

Participants in the *computer science app* study and *reading comprehension app* study showed a preference for learning with iPad applications compared to traditional or other methods, and in the case of the computer science app study the results were statistically significant for 4 out of the 6 applications tested. Participants in the *computer science app* study, *literacy app* study and *reading comprehension app* study suggested that future learners use the iPad application in addition to traditional and or other methods, and in the case of the computer science app study and the reading comprehension app study, the results are statistically significant. The results of the usability and user experience surveys, post-experiment questionnaire, and informal discussions with participants in the computer science app study, literacy app study and reading comprehension app study all support that gamification design approaches to tablet software lead to engagement.

The results of the quiz and practice sheets in the computer science app study and reading comprehension study support that the *software was educationally effective* in that students were able to learn using the software. Whether the software was *more* educationally effective than other methods

was not a research objective, and we cannot claim that our results suggest this.

As discussed when we introduced our hypothesis, we anticipated that some design approaches would be more effective than others, and part of our objective for conducting these studies was to determine which design approaches are more or less effective, in what context, and why, and, in the next section, we synthesize the commentary provided in each paper and discuss common findings and themes which emerged organically across the studies.

6.2 Analysis

An analysis of the results across the studies leads us to comment on the role of educational software, and the implementation and testing of software built using gamification design approaches.

6.2.1 Role of Educational Software

The user feedback in both the computer science app study and the literacy app study suggested that iPad applications were best suited for practicing the application of concepts, but that traditional instructional methods were better suited for teaching the concepts themselves. Students taking a first year computer science course and clients of an adult literacy centre are very different groups of people in terms of demographics, life experiences end educational experiences. The problem domains of computer science and literacy are also fairly different. The fact that this finding appeared so strongly in *both* groups and problem domains suggests that it is a deeper finding that will appear in other groups and problem domains as well. This finding is in-line with the historical role of video games in the classroom[Squ03]. Though the reading comprehension app study was designed to go beyond this role and teach concepts independently, participants still almost unanimously recommended that the iPad application and other instructional methods for future learners. For these reasons, we strongly suggest educational software should be first considered as a tool to augment and enhance existing instructional methods in learning environments, particularly as a method of practicing the application of concepts.

Despite the fact that our experiments were not designed or motivated by social interaction among participants, we observed such effects, and commented on them in the literacy app study. Participants began to both compete with one another and help one another during the literacy app study. Par-

ticipants who competed against one another benefited from the motivation provided, and participants that helped one another benefited from having the application as a reference point to discuss the problem and assist in understanding. Though we did not have space in the conference paper for the computer science app study to comment on this, we noticed the same competitive and helpful behaviours among participants in this study as well. The effect was a spontaneous instance of pair programming or pair learning[WK+00], facilitated by the ease with which students could pass around iPad devices. Again, the vast differences in participant demographics and problem domains leads us to conjecture that this is a widespread phenomenon, meriting further study to confirm these observations and to develop best practices. At this point, we suggest that educators consider using educational tablet software in environments conducive to socialization and safe device sharing (e.g. at a group table inside a classroom), and allow students to socialize while using the software.

6.2.2 Software Development

Several of the experimental applications were in whole or in part poorly received, although it is likely that these issues can be corrected.

There are many models and methodologies in software development, including the waterfall model, the spiral model and the agile development methodology[Boe88, CLC03]. In the waterfall model, the requirements for the software are first defined, the software is designed, implementated, tested and maintained. Vastly oversimplifying and ignoring important distinctions, the spiral model and the agile development methodology are both iterative cyclical versions of the waterfall model, involving iterative cycles of development incorporating different types of feedback.

In our view, software development methodologies such as the spiral model and the agile software methodologies are much better suited to the development of software utilizing gamification design approaches. Gamification is meant to encourage engagement, which is a subjective design requirement, dependent upon the emotional state of the software's users. Though best practices may be followed to increase the likelihood of a good initial design, ultimately the software must be tested with users to check for their level of engagement. The feedback from the users can then be used to fine tune the gamification design elements. Our studies should in some sense be viewed as the first iteration of what should ultimately be a cyclical, iterative development process that incorporates user feedback to design better software. For these reasons, we strongly suggest that developers of software incorporat-

ing gamification design elements choose a more cyclical software development methodology that allows for iterative re-designs, and account for the time user feedback takes to collect and to incorporate. Due to resource constraints, of all the experimental apps, only the binary numbers application has gone through this process, and is now available as "Image 2 Bits" for iPads.

Within an iterative framework, we suggest that discount usability testing be used early in the software development process[Nie94]. Discount usability testing is an adaptive testing approach that involves testing an application with as few as 3-6 participants for the purpose of informing design decisions. We recommend this approach because in all cases where participants pointed out problems with the design of the applications, 3-6 test participants and their informal feedback would have been enough to identify the issues. A more thorough usability study with more participants and more sophisticated instruments could then be done after improvements to the design had been made (perhaps after several more design iterations and discount usability tests). Our recommendation to use quick and dirty testing early in the development process followed by more extensive testing later has been suggested by others[SB10].

6.2.3 Gamification Design

Across the studies, we were able to confirm the effectiveness of known gamification design elements such as rewards, levels, and short-, medium- and long-term goals. Their implementation and fine tuning of course matters, but as a general rule these design elements can work in the domain of mobile educational software.

In both the computer science app study and the literacy app study, post-study questionnaire comments and informal feedback included particular praise for corrective feedback features. When users of these applications would make some mistakes, these applications would help the user see the correct answer through corrective feedback (e.g. an animation of the correct action in the CPU app). Corrective feedback is a known technique in both gamification and education literature[CCD12, BYC05], but the strength of strength and breadth of user comments merits highlighting.

6.3 Future Work

In addition to the future work suggested in each individual study, we suggest the following avenues for future work.

One unexpected observation in the computer science app study and the literacy app study was the impact of unplanned social effects. Due to their size and shape, making it easy to share, tilt and point to content, tablets encourage socialization. Studies investigating the relative impact of designed and impromptu social effects of mobile educational software versus other methods would help elucidate this. For example, testing applications conducive to sharing and/or competition with participants in isolation and in groups of varying sizes, or testing applications conducive to sharing on tablet devices versus desktop computers found in many institutional computer labs.

The time it takes to design, build and test a meaningful tablet application is considerable. The testing phase, in particular, involves soliciting participants, and collecting the data in time-consuming experiment sessions. A platform that would automate recruitment and allow usability researchers to deploy applications for remote testing has the potential to provide significant efficiency gains. For example, a platform that would help recruit study participants, allow study participants to download applications to their device, verify consent, conduct questionnaires on the device, and record and collate test results could automate many clerical tasks. Participants could even be given a reward (e.g. gift certificate) via the platform. Such a platform would be of interest to many usability researchers.

Bibliography

[A+10] David Autor et al. The polarization of job opportunities in the us labor market: Implications for employment and earnings. *Center for American Progress and The Hamilton Project*, 2010.

[Ada13] Ernest Adams. *Fundamentals of game design*. Pearson Education, 2013.

[Bar12] Frank Barry. How online fundraising, gamification and social media helped raise over $2 million dollars in one day, 2012. http://www.npengage.com/onlinefundraising/howonlinefundraisinggamificationandsocialmediahelpedraiseover2milliondollarsoneday/ (accessed April 5, 2013).

[BB85] Carl Bereiter and Marlene Bird. Use of thinking aloud in identification and teaching of reading comprehension strategies. *Cognition and instruction*, 2(2):131–156, 1985.

[BKM08] Aaron Bangor, Philip T. Kortum, and James T. Miller. An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6):574–594, 2008.

[BL91] Nanci Bell and Phyllis Lindamood. *Visualizing and verbalizing: For language comprehension and thinking*. Academy of Reading Publications Paso Robles, CA, 1991.

[Boe88] Barry W Boehm. A spiral model of software development and enhancement. *Computer*, 21(5):61–72, 1988.

[BS84] Thomas W Bean and Fern L Steenwyk. The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. *Journal of Literacy Research*, 16(4):297–306, 1984.

[BTNP05] Lynn Barr-Telford, François Nault, and Jean Pignal. Building on our competencies: Canadian results of the international adult literacy and skills survey. *Statistics Canada. Available at: www. statcan. ca/bsolc/english/bsolc*, 2005.

[BYC05] John Bitchener, Stuart Young, and Denise Cameron. The effect of different types of corrective feedback on esl student writing. *Journal of second language writing*, 14(3):191–205, 2005.

[CCD12] Frederik Cornillie, Geraldine Clarebout, and Piet Desmet. Between learning and playing? exploring learners perceptions of corrective feedback in an immersive game for english pragmatics. *ReCALL*, 24(03):257–278, 2012.

[CG11] Alma L Culén and Andrea Gasparini. ipad: a new classroom technology? a report from two pilot studies. *INFuture Proceedings*, pages 199–208, 2011.

[Che12] Brian X. Chen. Hmh fuse pilot program, 2012. http://www.hmheducation.com/fuse/pilot-1.php.

[Chi10] Oliver Chiang. Mark zuckerberg: ipad is not mobile, 2010. http://www.forbes.com/sites/oliverchiang/2010/11/03/mark-zuckerberg-says-ipad-is-not-mobile/ (accessed March 26, 2016).

[Cho03] Paul Chomsky. Why gizmos work: Empirical evidence for the instructional effectiveness of explorelearnings interactive content. `http://staff.explorelearning.com/pfaff/info/WhyGizmosWork.pdf`, 2003. Last accessed June 1, 2011.

[CK04] Dongseong Choi and Jinwoo Kim. 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, 2004.

[Cla87] C Clark. Abt, serious games, 1987.

[CLC03] David Cohen, Mikael Lindvall, and Patricia Costa. Agile software development. *DACS SOAR Report*, (11), 2003.

[Coh83] Ruth Cohen. Self-generated questions as an aid to reading comprehension. *The Reading Teacher*, pages 770–775, 1983.

[Cos08] Greg Costikyan. I have no words & i must design. 1994, 2008.

[DDKN11] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, MindTrek '11, pages 9–15, New York, NY, USA, 2011. ACM.

[Det11] S. Deterding. Situated motivational affordances of game elements: A conceptual model., 2011. Presented at Gamification: Using Game Design Elements in Non-Gaming Contexts, a workshop at CHI 2011. Retrieved from http://gamification-research.org/wp-content/uploads/2011/04/09-Deterding.pdf.

[DMT05] Richard Desjardins, TS Murray, and AC Tuijnman. *Learning a living: First results of the Adult Literacy and Life Skills survey*. OECD, 2005.

[Fri10] Stephen Fried. *Mobile Device Security: A Comprehensive Guide to Securing Your Information in a Moving World*. Auerbach Publications, 2010.

[Gar12] Gartner. Gartner says by 2014, 80 percent of current gamified applications will fail to meet business objectives primarily due to poor design, 2012. http://www.gartner.com/newsroom/id/2251015 (accessed April 5, 2013).

[Gar15] Gartner. Gartner says tablet sales continue to be slow in 2015, 2015. http://www.gartner.com/newsroom/id/2954317 (accessed May 4, 2015).

[GB86] Linda B Gambrell and Ruby J Bales. Mental imagery and the comprehension-monitoring performance of fourth-and fifth-grade poor readers. *Reading Research Quarterly*, pages 454–464, 1986.

[Gen90] James W Gentry. What is experiential learning. *Guide to business gaming and experiential learning*, pages 9–20, 1990.

[HBC+92] Thomas T Hewett, Ronald Baecker, Stuart Card, Tom Carey, Jean Gasen, Marilyn Mantei, Gary Perlman, Gary Strong, and William Verplank. *ACM SIGCHI curricula for human-computer interaction*. ACM, 1992.

[HE11] Juho Hamari and Veikko Eranti. Framework for designing and evaluating game achievements. *Proc. DiGRA 2011: Think Design Play*, 115(115):122–134, 2011.

[HH12] Kai Huotari and Juho Hamari. Defining gamification: a service marketing perspective. In *Proceeding of the 16th International Academic MindTrek Conference*, pages 17–22. ACM, 2012.

[Hig08] Reach Higher. America: Overcoming the crisis in the us workforce, 2008.

[HKS14] Juho Hamari, Jonna Koivisto, and Harri Sarsa. Does gamification work?–a literature review of empirical studies on gamification. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 3025–3034. IEEE, 2014.

[HL04] Chin-Lung Hsu and Hsi-Peng Lu. Why do people play on-line games? an extended tam with social influences and flow experience. *Information & Management*, 41(7):853–868, 2004.

[HY12] S. Henderson and J. Yeow. ipad in education: A case study of ipad adoption and use in a primary school. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 78–87, 2012.

[Kal15] Filiz Kalelioğlu. A new way of teaching programming skills to k-12 students: Code. org. *Computers in Human Behavior*, 52:200–210, 2015.

[Kap12] Karl M Kapp. *The gamification of learning and instruction: game-based methods and strategies for training and education.* John Wiley & Sons, 2012.

[KGB05] M Kutner, E Greenberg, and J Baer. National assessment of adult literacy (naal): A first look at the literacy of americas adults in the 21st century (report no. nces 2006–470). *Washington, DC: National Center for Education Statistics, US Department of Education*, 2005.

[Kol85] David Kolb. Learning styles inventory. *The Power of the 2 2 Matrix*, page 267, 1985.

[Leo13] Devin Leonard. The ipad goes to school, 2013. http://www.bloomberg.com/ bw/articles/2013-10-24/ the-ipad-goes-to-school-the-rise-of-educational-tablets (accessed August 4, 2015).

[Lit01] BC Literacy. *Who Wants to Learn?: Patterns of Participation in Canadian Literacy and Upgrading Programs: Summary of Results from the National Study Conducted for ABC Canada in Partnership with Literacy BC.* ABC Canada, 2001.

[Mac92] I. Scott MacKenzie. Fitts' law as a research and design tool in human-computer interaction. *Hum.-Comput. Interact.*, 7:91–139, March 1992.

[Mar98] Robert J Marzano. A theory-based meta-analysis of research on instruction. *Educational Research*, 80014(December):174, 1998.

[Mar01] K Maroney. My entire waking life. the games journal. *Retrieved Sept*, 11:2014, 2001.

[MC05] David R Michael and Sandra L Chen. *Serious games: Games that educate, train, and inform*. Muska & Lipman/Premier-Trade, 2005.

[McC12] Simon McCallum. Gamification and serious games for personalized health. *Stud Health Technol Inform*, 177:85–96, 2012.

[Nic12] Scott Nicholson. A user-centered theoretical framework for meaningful gamification. *Games+ Learning+ Society*, 8(1), 2012.

[Nie94] Jakob Nielsen. Guerrilla hci: Using discount usability engineering to penetrate the intimidation barrier. *Cost-justifying usability*, pages 245–272, 1994.

[Pec13] Teri Pecoskie. Public school board rolling out expansive ipad project, 2013. http://www.thespec.com/news-story/ 4052858-public-school-board-rolling-out-expansive-ipad-project/ (accessed August 4, 2015).

[REZ03] Stephen P. Rushton, Janice Eitelgeorge, and Ruby Zickafoose. Connecting brian cambourne's conditions of learning theory to brain/mind principles: Implications for early childhood educators. *Early Childhood Education Journal*, 31:11–21, 2003. 10.1023/A:1025128600850.

[RMR10] Paula Rego, Pedro Miguel Moreira, and Luis Paulo Reis. Serious games for rehabilitation: A survey and a classification towards a taxonomy. In *Information Systems and Technologies (CISTI), 2010 5th Iberian Conference on*, pages 1–6. IEEE, 2010.

[Ros97] Barak Rosenshine. The case for explicit, teacher-led, cognitive strategy instruction. *MF Graves (Chair), What sort of comprehension strategy instruction should schools provide*, 1997.

[Ruf13] Cory Ruf. Steel shutdown: The decline of hamilton's manufacturing, 2013. http://www.cbc.ca/news/ canada/hamilton/news/ steel-shutdown-the-decline-of-hamilton-s-manufacturing-1.2350633 (accessed August 4, 2015).

[SB10] Ben Shneiderman and Shneiderman Ben. *Designing The User Interface: Strategies for Effective Human-Computer Interaction, 5/e (New Edition)*. Pearson Higher Education, 2010.

[SCC10] Hui-Fang Shang and I-Ju Chang-Chien. The effect of self-questioning strategy on efl learners reading comprehension development. *The International Journal of Learning*, 17(2):41–54, 2010.

[Sha91] Brian Shackel. Usability-context, framework, definition, design and evaluation. *Human factors for informatics usability*, pages 21–37, 1991.

[She14] Esther Shein. Should everybody learn to code? *Communications of the ACM*, 57(2):16–18, 2014.

[Shn83] B. Shneiderman. Direct manipulation: A step beyond programming languages. *Computer*, 16(8):57–69, 1983.

[SJB07] Tarja Susi, Mikael Johannesson, and Per Backlund. Serious games: An overview. 2007.

[SSGRB02] Wilmar B Schaufeli, Marisa Salanova, Vicente González-Romá, and Arnold B Bakker. The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. *Journal of Happiness studies*, 3(1):71–92, 2002.

[SZ04] Katie Salen and Eric Zimmerman. *Rules of play: Game design fundamentals*. MIT press, 2004.

[Tou11] Kayne Toukonn. The dynamic electronic textbook: Enhancing the student's learning experience. *Kent State University*, 2011. MFA Thesis.

[WCD11] K. Wattanatchariya, S. Chuchuaikam, and N. Dejdumrong. An educational game for learning wind and gravity theory on ios: Drop donuts. In *Computer Science and Software Engineering (JCSSE), 2011 Eighth International Joint Conference on*, pages 387 –392, may 2011.

[WH97] Jane Webster and Hayes Ho. Audience engagement in multimedia presentations. *ACM SIGMIS Database*, 28(2):63–77, 1997.

[WJ82] Bernice YL Wong and Wayne Jones. Increasing metacomprehension in learning disabled and normally achieving students through self-questioning training. *Learning Disability Quarterly*, 5(3):228–240, 1982.

[WK+00] Laurie Williams, Robert R Kessler, et al. The effects of" pair-pressure" and" pair-learning" on software engineering education. In *Software Engineering Education & Training, 2000. Proceedings. 13th Conference on*, pages 59–65. IEEE, 2000.

[Yan11] Feng Yan. A sunny day: Ann and rons world an ipad application for children with autism. In Minhua Ma, Manuel Fradinho Oliveira, and Joo Madeiras Pereira, editors, *Serious Games Development and Applications*, volume 6944 of *Lecture Notes in Computer Science*, pages 129–138. Springer Berlin / Heidelberg, 2011.

[Zam12] Zamzee. New research shows zamzee increases physical activity by almost 60%, 2012. http://blog.zamzee.com/2012/09/26/newresearchshowszamzeeincreasesphysicalactivitybyalmost60/ (accessed April 5, 2013).