Algorithmic Generation and Mobile Distribution of Phonetic, Orthographic, and Inference-Based Literacy Exercises for Adult Learners

by Jennifer Rose Hill

B.S. in Computer Science, May 2012, Hood College

A Dissertation submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

January 19, 2018

Dissertation directed by

Rahul Simha
Professor of Computer Science

The School of Engineering and Applied Science of The George Washington University certifies that Jennifer Rose Hill has passed the Final Examination for the degree of Doctor of Philosophy as of December 13, 2017. This is the final and approved form of the dissertation.

Automatic Generation and Mobile Distribution of Phonetic, Orthographic, and Inference-Based Literacy Exercises for Adult Learners

Jennifer Rose Hill

Dissertation Research Committee:

Rahul Simha, Professor of Computer Science, Dissertation Director

Rachelle Heller, Professor of Computer Science, Committee Member

Terry Salinger, Chief Scientist for Reading Research, Workforce and Life Long Learning Program, American Institutes for Research, Committee Member

Abstract of Dissertation

Algorithmic Generation and Mobile Distribution of Phonetic, Orthographic, and Inference-Based Literacy Exercises for Adult Learners

A 2014 study found that more than half of all Americans aged 16-74 do not possess the literacy skills needed to adequately cope with the demands of day-to-day life, with over 13 million people in the United States alone considered to be functionally illiterate. There have been numerous efforts to address these high illiteracy rates by way of Adult Basic Education programs; however, such programs are consistently underfunded and plagued with high drop-out rates, as many students struggle to consistently attend classes due to lack of transportation, uncontrollable family circumstances, or conflicting employment hours. At the same time, smartphones are becoming ever more ubiquitous, signifying an opportunity for mobile educational software to address many of these difficulties.

This thesis describes the creation of a software system to allow students in such programs to continue their learning outside the classroom. This project, known as CAPITAL (Comprehension and Pronunciation Instructional Tools for Adult Learners), involves three core components. The first is a system for algorithmically generating questions to thoroughly cover the different microskills of alphabetic literacy, the beginner level of reading. This system is capable of generating materials for testing every facet of phonological awareness, phonemic awareness, decoding, and encoding, customized to accommodate any desired skill progression. Second is a system for algorithmically generating questions for reading comprehension at the intermediate reading level that specifically target the reader's ability to draw inferences and monitor their own

comprehension. The final component is a mobile-learning application, carefully designed for ease of use by adults with below-average literacy, through which students receive these generated learning materials. Trial studies were conducted with literacy providers in the Washington D.C. area to seek feedback from both teachers and students.

The combination of these three components results in a mobile application that is easy for adult learners to use, provides them with a substantial amount of high-quality learning materials, seamlessly integrates with the lessons they are learning in their classes, and delivers materials to them at an optimal and customized rate. CAPITAL is the first system of its kind to address the adult literacy crisis by both allowing for scalable automated creation of learning materials and providing a more effective method of delivering them to students in need.

Table of Contents

List of Figures

List of Tables

Glossary of Terms

Phoneme: The smallest unit of sound in a language

Grapheme: The smallest meaningful written unit in a language

Orthography: The letters and written patterns that make up the sounds of a language

Chapter 1 - Introduction

1.1 The Problem

In 2012, the Organization for Economic Cooperation and Development (OECD) conducted a large-scale international study to evaluate and compare the general skills of adults around the world, of which literacy was a major focus. This initiative, referred to as the Program for the International Assessment of Adult Competencies (PIAAC), measured literacy proficiency using a scoring metric adapted from the earlier International Adult Literacy Survey (IALS) [1], grouping numerical scores into broader skill levels ranging from 1 to 5. The IALS considers a proficiency level of 3 to be the minimum skill level necessary for an adult to be able to adequately cope with the demands of day-to-day life, which corresponds to roughly the ability level of a high school graduate.

However, a 2014 U.S. household study of 8,700 adults conducted by the PIAAC found that 51% of Americans aged 16-74 failed to reach this minimum benchmark of adequate literacy, exhibiting proficiencies at a level 2 or below [2]. Approximately 4% of the population did not even reach level 1 proficiency, making them unable to perform even the simplest of literacy tasks. These individuals are what is known as *functionally illiterate*: they are unable to "engage in all those activities in which literacy is required for effective functioning of [their] group and community" [3].

By current U.S. population numbers, this statistic represents nearly 13 million people who lack even the most basic of reading skills, and nearly 165 million people who struggle to read at an adequate enough level to get by day-to-day.

1.2 What is Literacy?

The PIAAC defines literacy as "understanding, evaluating, using, and engaging with written texts to participate in society, to achieve one's goals, and to develop one's knowledge and potential" [4]. However, it is difficult to encompass all the nuances of the term within a single sentence definition. In a narrow sense, literacy can simply refer to one's ability to translate written words into their verbal sounds, a process known as *decoding*. In a broader sense, however, literacy refers to comprehension, the ability to construct meaning from and draw inferences about the information being read.

Literacy encompasses a wide spectrum of skills, beginning with the ability to sound out a single written letter and culminating in a deep and thorough understanding of the meaning behind written text. Additionally, there are three distinct categories of literacy according to the National Assessment of Adult Literacy, each of which encompasses its own set of skills [5]. Following these standards, a person who is fully literate must possess *prose literacy*, the ability to navigate and comprehend continuous texts such as news articles and books; *document literacy*, the ability to navigate and comprehend non-continuous texts such as job applications and maps; and *quantitative literacy*, the ability to perform computations using numbers in printed materials, such as balancing a checkbook or calculating a tip. A deficiency in any one of these forms of literacy can be detrimental to a person's quality of life.

1.3 How Literacy is Acquired

When it comes to *acquiring* literacy, the National Adult Literacy Survey (NALS) employs a two-part definition: the first entails learning how one's language is encoded in its respective writing system, while the second refers to the ability to apply this

knowledge to literacy situations [5]. This definition assumes that the individual already has knowledge of the language itself in its spoken form, including its syntax and at least a general functional knowledge of its vocabulary.

The end goal of literacy acquisition is the construction of meaning: in other words, comprehension [6]. Reading is a multifaceted skill that seamlessly and automatically integrates attention, memory, language, and motivation [7]. As a reader's skills progress, their focus automatically shifts from simply identifying the words on a page to obtaining a deeper understanding of the meaning of the language and the intentions behind it.

Although most of the studies pertaining to literacy acquisition have focused on children, research has shown that adults are just as capable of acquiring and improving their literacy skills, regardless of age. The process typically takes longer for adults than for children [8], but despite differences in cognitive abilities between younger and older learners, the efficiency and effectiveness of learning to read is largely the same for learners of any age [9]. Adults who are learning to read have even been shown to follow the same general stages of literacy acquisition as do children; in other words, the same types of failures that children encounter when first learning to read are also demonstrated by adults with poor reading skills [10, 11, 12].

The past several decades of research have settled on three primary skills that comprise the core components of reading, which are covered by most literacy education programs today [7, 13, 14]. Each of these skills can be viewed as a discrete stage of reading development, where each stage is dependent on the concepts within the previous stages and must be mastered before the next stage is reached [15]. These topics are:

1. **Alphabetics**: knowledge of the sounds that make up a language and the ability to connect these sounds to their written representations;
2. **Fluency**: the ability to read quickly and accurately, with the appropriate rhythm, intonation, and expression;
3. **Comprehension**: the ability to construct meaning from written text

Each of these topics is described in more detail in the following sections.

1.3.1 Alphabetics

Alphabetics refers to the process of deriving meaningful spoken words from the written letters in the alphabet [16]. Alphabetic proficiency manifests in two distinct ways: phonological awareness, the knowledge of the sounds that make up spoken language; and word analysis, the knowledge of the relationship between written letters and the sounds they represent. Each of these skills is described in the following sections.

1.3.1.1 Phonological Awareness

Spoken dialogue is made up of individual words strung together in a chain, and each of these words is made up of even smaller units of sound; the word *begin*, for example, can be broken up into syllables (*be-gin*), or an onset (*b*) and rime (*egin*), or into phonemes, its smallest sound units (b-ɪ-g-ɪ-n). Phonological awareness, the innate understanding that spoken language is made up of smaller units of sound and the ability to isolate and manipulate these sounds, is an essential foundational skill for reading.

Phonological awareness exists as a continuum of complexity, from the most basic ability to segment sentences into their component words, to identifying rhyming sounds and alliteration, and advancing to an awareness of syllables, onset and rimes, and finally

phonemes [17, 18, 19]. Awareness of phonemes is considered to be the most advanced stage of phonological awareness and is essential for mastering word-reading skills [20].

Although instruction in the simpler phonological components like rhymes and syllables has been shown to be a prerequisite for the development of more complex sound processing skills [21], the development of these skills alone has shown no direct benefit to the reading process [22]. The true keystone of reading acquisition at the word level is *phonemic* awareness, the ability to manipulate words at the phoneme level.

In the majority of cases, individuals who struggle with basic word reading tasks are hindered by a lack of phonemic awareness, leading to an inability to process language phonetically [23, 24]. [25] posits that this pattern holds true because an inability to adequately process phonemes interferes with a person's ability to build and maintain verbal representations of what they are reading in their short-term memory; strengthening this skill, therefore, would also improve a reader's ability to better access this memory mechanism.

Numerous studies have demonstrated that phonological and phonemic awareness are critical components of reading success, the vast majority of which have focused on the developing literacy of children. [26] identified a strong correlation between lack of phonological processing skills and poor reading abilities in children of varying ages, determining that a child's awareness of phonological structure early in life serves as a dependable predictor of future reading success. [27] found that first graders who were unable to do certain phonemic manipulations (e.g. blending, segmenting, and replacing sounds) remained in the bottom quarter of their class for reading ability even as they reached fifth grade. [28] observed significant improvement in the reading and spelling

skills of kindergarten children whose lessons included phoneme awareness training as opposed to just learning letter/sound relationships. Even children with limited knowledge of spoken English follow this same pattern: students who had developed phonological awareness skills in their native language were found to be more successful at learning to read in English as a second language [29].

Several studies have confirmed that this same pattern holds true for low-literate adults, irrespective of their education levels or general cognitive maturity. [11] compared the reading skills and phonemic awareness of low-literate adults in prison, finding that they experienced the same difficulties in decoding and segmentation as did children with poor reading ability, despite exhibiting no symptoms of intellectual deficiency. [30] confirmed that this same correlation between lack of phonological awareness and poor reading ability also exists in otherwise educated adults and for those enrolled in basic education programs. Some of the adults who were classified as poor readers were high school graduates and had been actively enrolled in literacy instructional programs for several years, demonstrating that deficits in phonemic processing skills can hinder the improvement of one's literacy even in the presence of other educational interventions. These studies provide evidence to confirm the hypothesis that awareness of phonemes does not develop organically over time, and must instead be explicitly fostered through deliberate practice and instruction in order for literacy skills to develop [31, 32, 33].

Intensive phonemic awareness instruction is particularly necessary for adults, as most adult learners, even those with below-basic literacy skills, can identify many words by sight alone, despite being unable to truly read them. For example, an illiterate adult who drives regularly might be able to identify the word "stop" due to frequent exposure to the

word on street signs, yet wholly unable to read similar words like “shop” or “stomp” [32]. Studies have shown that adults actually have more success when reading connected texts as opposed to individual words, because their reading strategy leverages sight word familiarity to provide context to allow them to guess at the unfamiliar words [34, 35]. Phonemic awareness is critical to bridging this skill gap and helping readers to un-learn these ultimately disruptive coping strategies.

1.3.1.2 Word Analysis

Phonological awareness exists entirely within the realm of sound, addressing a reader’s ability to identify and manipulate the smaller sound units that make up spoken language. While this skill is a prerequisite to successful reading, it does not actually incorporate written language, without which reading is impossible. The transition from sound-sound to word-sound relationships manifests as two distinct skills, known as *decoding* and *encoding*. The term decoding refers to the ability to parse a written word and translate it into its pronunciation, while encoding refers to the ability to translate a spoken word into its written form. More colloquially, these two terms can be identified as “reading” and “writing,” respectively.

Studies have shown that weak decoding skills are the most common cause of reading comprehension failure; proficiency in reading at the word level is mandatory, although not sufficient, for successful comprehension [32, 36]. A skilled reader can accurately decode a given word in an isolated context using a combination of strategies, including identifying its visual pattern, blending its component graphemes phonetically, or using parts of familiar words to inform the word’s makeup [37].

The process of encoding spoken words into their written forms is an equally central component in the development of word knowledge; low-literate adult readers tend to make frequent spelling errors, many of which demonstrate a lack of awareness of orthographic patterns and their connections to phonetic sequences [12]. The processes of decoding and encoding have been proven to be intrinsically linked, drawing upon the same underlying knowledge of the relationship between the orthographic and phonological makeup of words [38, 39]. As such, the skills of phonemic awareness, decoding, and encoding are all interconnected, and proficiency in each of these skills is necessary for successful comprehension.

1.3.2 Fluency

Researchers used to believe that decoding failure was the primary "bottleneck" for poor reading; it was assumed that once a reader was able to identify every word in a passage, comprehension would follow automatically [36]. However, more recent research has proven that decoding skills are not the only barrier to reading comprehension: in fact, strengthening decoding skills alone has been shown to have little to no impact on comprehension ability [40, 41]. True comprehension cannot be obtained unless decoding becomes a fast and automatic process, in a process known as *fluency*. The reason for this can be attributed to the theory of automatic information processing [42]: readers need to be able to perform surface-level processing of text without exhausting their cognitive capacity. Because humans have finite mental resources and limited attention spans, the more effort that a reader needs to devote to decoding words, the less energy they will have available for comprehending their meaning [15, 38].

There are three unique components that contribute to reading fluency: rate, accuracy, and prosody. Rate refers to the speed at which words are read; if the decoding process takes too long, the decoded material is likely to be forgotten before it can be processed [43]. However, accuracy is also a critical component of the reading process, as mistakes in decoding inherently interfere with the ability to fully understand the text. The third component, prosody, refers to reading with the proper expression, addressing intonation, stress, rhythm, and rate when verbalizing a written text [15].

Fluency can be viewed as the "bridge" between decoding and comprehension [44]. It integrates the alphabetic principle and the understanding of phoneme-grapheme relationships into a fast and familiar process, allowing readers to transition from reading letters and words in isolation to reading phrases, sentences, and beyond. Substantial research has shown that fluency is required for reading comprehension [15, 40, 45].

1.3.3 Comprehension

Comprehension, the ultimate goal of reading, is a multi-faceted process that employs a wide array of different mental skills. To comprehend a text, readers must be able to identify and summarize its main ideas, draw inferences, ask and answer questions, and synthesize all of this information into a cohesive mental representation that integrates with prior background knowledge, in what is known as a *mental model* [46] or *situation model* [47]. To build a successful and accurate mental model of a text, the reader must be able to draw on the foundational skills discussed in the prior sections. A reader must be able to decode the words that make up the text they are reading with enough speed and automaticity to leave most of their mental resources free for comprehension.

With a well-formed mental model, readers can access information from the text they have read to recall facts and compare information to their established background knowledge, allowing them to build greater networks of understanding. A mental model allows a reader to encode meaningful information about a text, including the causal relationship between events that take place, spatial and temporal information, and actors and their actions [48].

Understandably, reading comprehension is significantly improved when the reader has knowledge of the meaning behind the words being read [49]. In fact, along with decoding deficits, a lack of vocabulary understanding is one of the two most common sources of comprehension failure [50]. Adults in the beginning stages of reading tend to have a more advanced vocabulary than children at the same reading level, due to their general world knowledge and greater exposure to spoken language [51]. However, as reading advances into the intermediate levels, printed language quickly surpasses speech as a method of furthering vocabulary knowledge, significantly hindering the vocabulary growth of non-readers [52, 53]. Successful comprehension of a text requires the reader to be familiar with only 90-95% of the words within, leaving the remaining words to be learned organically without the need for external intervention [54].

1.4 Adult Basic Education Programs

The previous section outlined the extensive body of research that has examined the ways in which human beings learn to read. These findings have been applied to classrooms around the world to better assist teachers in their quest to instill strong reading skills in children. Yet despite this, the United States continues to see staggering numbers of adults with less than sufficient reading ability.

There have been numerous efforts to address the high rate of functional illiteracy in U.S. adults by way of Adult Basic Education (ABE) programs. Many such programs exist which are designed to help adults improve their literacy skills in traditional classroom settings. However, these programs do not nearly have the resources necessary to reach all the adults who need them. As of 2015, only 4.1 million adults were enrolled in adult education programs, encompassing only 11% of the low-literate adults who could benefit from them [55]. This is in large part due to the inability of these programs to support greater numbers of students: 67% of programs today reported having more students wanting to attend than can be supported by the current program resources [56].

Adult education programs are also consistently underfunded. In 2008, only 49% of ABE programs were receiving some form of federal or state funding, and as of 2016, this number had dropped to 36% [56]. With such little funding to go around, many programs struggle to stay open and to employ a sufficient number of teachers, making it even harder for learners to find a program that can support them for long enough to help them make any actual learning progress.

Even those adults who do participate in basic education are rarely able to benefit from these programs to the fullest extent, primarily due to their inability to stay in a program for long enough to make significant learning gains. ABE programs across the country are consistently plagued by high dropout rates, with estimates ranging from 60% to 80% student attrition on average [57, 58]. A 1994 survey by the National Evaluation of Adult Education Programs (NEAEP) found that half of all adults who enroll in an adult education program drop out before completing 35 hours of instruction, and only 11% of students remain in a program for a year or longer [59]. Studies have shown that adults

need to commit roughly 120 hours of learning to improve their reading by the equivalent of a single grade level [60], meaning that the vast majority of these adults leave their programs before realizing any significant improvements in their reading skills.

However, these bleak numbers do not suggest that adult basic education is not useful, nor does it necessarily imply a lack of student motivation or desire. The NEAEP survey found that 45% of the students who dropped out early in their programs did so because of external factors such as lack of transportation, uncontrollable family circumstances, or employment changes [59]. Another survey found that 73% of those students who dropped out of their respective programs reported that they would be willing to return under different conditions [61]. These numbers suggest that there is a great opportunity to improve the way that adult students experience ABE programs by making it easier for them to reap the benefits of structured education without being hindered by limited resources and the demands of physical attendance.

1.5 Potential for Technology

One method that these programs can use to foster continued motivation in students, reduce the amount of time needed to attend physical classes, and alleviate many of the difficulties associated with physical attendance is to provide learners with practice materials for use outside of the classroom. Extending the scope of learning to outside the classroom allows students to practice on their own schedule and better maintain what they have already learned. Students who currently drop out of programs due to external factors such as scheduling conflicts and transportation difficulties would be more able to keep up with the in-class curriculum without falling behind, decreasing the need for them to drop out and giving them greater control over their learning opportunities.

While workbooks and written practice materials are a common resource given to students for home practice, there are many benefits to providing these materials through software instead. Software systems can give users access to a larger number of learning materials at one time, and can guide them through materials in a structured and organized environment. Software learning tools are also dynamic: students can receive real-time feedback on their performance to help them better assess their progress, and the system can adapt to the individual user's strengths and weaknesses to provide each user with customized instruction. These factors allow learning software to provide a more personalized, convenient, and accessible learning experience than traditional pen-and-paper methods.

In the past decade, billions of dollars have been invested into the development of innovative educational technology for K-12, postsecondary, and corporate learning environments; however, very little money is being directed towards applying these same innovations to adult education [62]. Yet more than 85% of adult education administrators and instructors reported that they see the potential for educational technology to support and improve the effectiveness of their programs, primarily due to its ability to attend to students' individual needs. Instructors pinpointed the following four features as the most desirable benefits to introducing technology into their programs: (1) providing students with practice outside of the classroom; (2) allowing students to advance at their own pace; (3) providing personalized instruction for each student; and (4) allowing instructors to monitor their students' progress and performance [62]. Additionally, the regular use of software tools would also allow students to develop greater digital literacy. One instructor noted that "most entry-level jobs now require a basic comfort with technology,

and students need to develop proficiency to communicate within society" [62]. With technology ever advancing, both traditional literacy and digital literacy are necessities for nearly every type of profession.

While any software solution can provide the aforementioned benefits to learners compared to traditional classroom learning, technologies designed for mobile platforms in particular possess a significant benefit over other software: the portable nature of smartphones and tablets allows mobile software to be used virtually anywhere and at any time. This provides learners with even more control over when and where they wish to learn, relieving them of the need to set aside time to sit down at a computer specifically to study. For adult learners in particular, whose busy lives frequently interfere with their ability to consistently attend classes, this is a particularly valuable feature.

Mobile technology is becoming an ever more viable method of reaching the general population. Smartphone and tablet ownership rates in the United States have been growing steadily every year since 2011: surveys conducted by the Pew Research Center show that, by the end of 2016, 77% of all adults in the United States owned a smartphone, compared to a mere 35% five years prior [63]. Although there have not yet been any studies that specifically look at the number of low-literate adults in the United States who own smartphones, a report by [62] estimated that between 55% and 75% of adults enrolled in adult education programs across the country owned smartphones as of 2015. Research conducted by the NALS in 2002 has also shown significant correlations between poor literacy skills and a variety of demographic and socioeconomic factors [64], from which we can extrapolate the potential accessibility of mobile applications for this population. Illiteracy has been found to be substantially more likely to occur in adults

from the lowest income and education brackets, and as of 2016, smartphone penetration had reached more than half of the populations of both demographics: 64% of individuals from low-income households, and 54% of individuals with less than a high school education [63]. The NALS also found a high correlation between low literacy skills and respondents of African American or Hispanic descent [64], both populations which had achieved over 70% smartphone penetration in 2016 [63].

Perhaps most importantly, these same demographic groups are also some of the most likely to be "smartphone dependent": that is, their smartphone serves as their most reliable—and often their *only*—way of accessing the internet. In comparison to the smartphone ownership numbers described above, only 50% of low-income individuals and 29% of individuals with less than a high school education reported even owning a desktop or laptop computer. These numbers indicate that smartphones are perhaps the most ideal platform for providing learning resources to individuals in these demographic groups, when compared to non-mobile software solutions.

Studies have also shown that the number of teenagers and young adults in adult education programs is growing [65], and a remarkable 92% of individuals in this demographic today are smartphone owners [63]. In fact, younger learners have been found to be more likely to drop out of adult education programs because they feel as though their learning styles and needs are not being adequately met: younger generations expect and desire technology integration in their education [66] [67]. All of this data suggests that smartphones have become an increasingly viable method, and perhaps even a superior alternative, for providing digital materials to the low-literacy population.

1.6 Proposed Solution

The purpose of this thesis is to address the severe lack of useful educational technology for low-literate adult learners by developing a software system that can be used to supplement classroom learning for adults in basic education programs.

Such a system involves two core components. The first component is a system for creating appropriate materials for these learners. For the application to be an effective supplemental learning tool, the materials that students receive through the software must echo what they are learning in their classes. The materials provided by the software should not compete or conflict with the curricula established by the class teachers; rather, the tool should be designed *for* instructors, to allow them to customize the materials to align with their classroom teachings. Rather than requiring instructors to hand-create materials for their students, a process which is exhausting and time-consuming, this system will automatically generate materials that conform to the parameters established by the instructor. The end goal of such a system is to serve as a tool for creating an exhaustive, accurate, and customized set of materials that maximize the coverage of all the different skills that an instructor wishes their students to learn, organized in a spectrum of increasing difficulty.

The second component is a smartphone application for students, through which they can receive these generated learning materials and monitor their progression over time. The materials that students receive should always be level-appropriate and customized to their individual needs, providing them with a dynamic and carefully-controlled learning environment. This app must be carefully designed to be intuitive and easily learnable for adults with below average literacy so that they can comfortably use it with minimal to no

supervision, and it must provide an enjoyable user experience for them rather than being viewed as an additional burden to the learning process.

The combination of these two components will result in a mobile application that is easy for adult learners to use, provides them with a large number of high-quality and instructor-approved learning materials, seamlessly integrates with the lessons they are learning in their classes, and delivers materials to them at an individualized rate based on their performance.

1.7 Contributions of This Thesis

The following chapters discuss the design and implementation of the proposed software system, known as *CAPITAL*: Comprehension and Pronunciation Instructional Tools for Adult Learners. This thesis presents three novel contributions to the current state of the art in both question generation and educational software design.

The first contribution is a system for generating questions to thoroughly cover the full array of micro-skills at the alphabetics level. This system is capable of generating materials to test every facet of phonological awareness, phonemic awareness, decoding, and encoding. To the best of our knowledge, this is the first work of its kind to address the challenge of automatic generation of learning materials at the level of phonological and alphabetic literacy, making this thesis the pioneering research in the space.

The second contribution will explore the automatic generation of several different types of questions that specifically target common difficulties faced by poor comprehenders: inference making and comprehension monitoring. This thesis will describe a unique method for finding the most contextually relevant words in a text and introducing deliberate inconsistencies in their place that require inferences and contextual

awareness to identify. Also described is a novel application of a discourse parser for creating questions that challenge a reader's application of discourse connectives, a known difficulty for poor comprehenders. To our knowledge, no previous question generation systems have sought to target either of these specific comprehension challenges.

The final contribution of this thesis is a detailed account of the design of a smartphone application that is intuitive and usable by adults with below average literacy. The rationale behind the design decisions and the usability test results will neatly contribute to the existing literature on both mobile learning system design and accessibility design for low-literate users, as this is the first study to address the specific challenges of uniting both concepts in a single application.

Chapter 2 - Generating Alphabetics Items

This chapter discusses the design of an automatic generation system for creating practice materials at the alphabetics level. The goal of this system is to achieve complete coverage of nearly every facet of alphabetics instruction, which can then be applied to any desired curriculum to effortlessly produce an exhaustive set of materials to target any alphabetics-level skill. With the items generated, this system aims to ensure complete mastery of any individual word in a given curriculum, from its pronunciation, to its spelling, to its phonetic and orthographic relationships to other words.

Section 2.1 describes the individual components that comprise the lexicon from which items are generated, and outlines a novel algorithm for deriving the proper alignment of letters within a given word and the phonemes that they correspond to. Section 2.2 describes how this lexicon is applied to find logical groupings of words according to the lexical relationships they share: rhyming (2.2.1), orthographic similarities (2.2.2), and phonetic similarities (2.2.3). Section 2.3 outlines a series of algorithms used to generate different types of misspellings for a given word. In Section 2.4, each of the previous-described generators is evaluated on the validity of their output (2.4.1) and their ability to cover the words within a given curriculum (2.4.2). Section 2.4.2 also details a method for constraining the output of the generators to align with the skill progression of a given curriculum, and the results of the constrained and unconstrained data sets are compared for coverage. Finally, Section 2.5 summarizes the contributions of the chapter and their implications for future work.

2.1 Building the Lexicon

To target phonological and phonemic awareness, the proposed algorithms must be able to generate items that target a user's ability to identify, distinguish between, and manipulate the component sounds of a word. To accomplish this, the system must have access to a representation of a given word's pronunciation as well as its spelling. These word features are compiled in a *lexicon* to inform the generation process.

The transcribed pronunciation of a given word is obtained using the Carnegie Mellon University Pronouncing Dictionary (CMUdict[1]). The CMUdict contains over 134,000 English words and a phonemic breakdown for each one, transcribed using a set of 25 consonant phonemes and 15 vowel phonemes (complete with lexical stress) from the ARPAbet speech recognition symbol set. See Appendix A for the complete list of phonemes in the ARPAbet symbol set and an example of each.

The lexicon also includes a representation of the syllable boundaries of a word, both orthographically and phonetically. Each written word's hyphenated form was obtained from the Wordnik[2] online dictionary, whereas the syllable boundaries of each word's phonemic breakdown were obtained from an augmented version of the CMUdict from the work of [68].

To allow for exercises that require the manipulation of written words in relation to their pronunciations, it is necessary for the system to understand the relationship between the letters in a word's spelling and the phonemes that correspond to them. The task of determining the phonetic representation of a word from its spelling is known as grapheme-to-phoneme conversion. This is not a trivial task, because although a word's

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[2] http://www.wordnik.com

spelling directly influences the way it is pronounced, there is not always a one-to-one relationship between the letters and phonemes that make up a word. A single grapheme does not always map to the same phonetic representation: *think* and *bathe*, for example, both contain the grapheme *th*, but the two instances map to the phonemes `/TH/` and `/DH/`, respectively. Complicating things further, a string of consecutive consonants or vowels may represent a single sound when they function as a diphthong (such as the "ea" in *beaver*), or each letter may be individually vocalized (such as the "ea" in *create*). Some letters and letter clusters are not even represented phonetically at all: the *e* in *make*, the *gh* in *night*, and the *c* in *scissors* are "silent" letters which by themselves do not map to any vocalized phoneme.

The CMUdict provides the correct phonetic breakdown for all dictionary words, meaning that we do not need to address the task of letter-to-phoneme conversion. However, even given a word's spelling and phonetic breakdown, there is no reliable way to determine which specific letters produce which phonemes, as the pronunciation dictionary provides only a sequence of letters and a sequence of phonemes for each word, but no relationship between the elements within these sequences. For simple words, the task is often trivial: the spelling *b-a-t* and the phonemes `/B/-/AE/-/T/` have an obvious one-to-one correspondence; however, the relationship between *c-h-a-l-k* and `/CH/-/AA/-/K/` does not follow this same rule, posing a greater challenge to derive.

Letter-to-phoneme alignment is typically performed using machine learning, where a model is trained to predict the proper alignment of letters and their phonemes from a training data set of hand-annotated correct alignments [69]. However, CAPITAL's lexicon allows for a method of achieving letter-phoneme alignments that does not require

such a training set. The following section describes the algorithm used by the CAPITAL system to align the letters in a given word to their component phonemes.

2.1.1 Mapping Letter and Phoneme Clusters

This section proposes a novel approach to solving the letter-phoneme mapping problem by examining the relationship between the written and phonetic syllables in a word. The Oxford English dictionary defines a syllable as "a unit of pronunciation having one vowel sound, with or without surrounding consonants" [70]. Because a syllable, by definition, must contain only one vowel sound, we can view any given syllable as a sum of three distinct components: leading consonants (or *onset*), a vowel, and trailing consonants (or *coda*). This holds true for both the written and phonetic syllabication of a word.

Algorithm 1 Algorithm for splitting a syllable into onset, vowel, and coda

```
clusters = []
for each syllable pair (L, P) in word:
    v1: index of vowel string
    V2: index of vowel phoneme
    vowel = (Lv1, Pv2)
    onset = (L0 … Lv1, P0 … Pv2)
    coda = Lv1+1 … Ln, Pv2+1 … Pn)

    clusters += (onset, vowel, coda)
```

We can assume that any set of sequential vowels within a single syllable must come together to form the single vowel phoneme in that syllable. (For this purpose, "w" and "y" are also included when appearing at the end of a vowel string, as both letters are semivowels that can function as vowels when they are not themselves vocalized. A trailing "r" is also included in the vowel string when the resulting vowel phoneme is */ER/*.) Consequently, any letters that come before or after the vowel string must serve as

the onset and coda, respectively, and can be mapped to the consonant phonemes before and after the vowel (see Algorithm 1). An example of the desired output can be seen in Figure 2-1.

	Onset	Vowel	Coda
flies	f l	i e	s
/F/ /L/ /AY/ /Z/	*/F/ /L/*	*/AY/*	*/Z/*

Figure 2-1. A word's letters and phonemes correctly split into their onset, vowel, and coda

Relying on syllabication also addresses the challenge of distinguishing between vocalized letters and diphthongs. For example, the word "beaver" is divided into *bea* (*/B/ /IY/*) and *ver* (*/V/ /ER/*), whereas "create" becomes *cre* (*/K/ /R/ /IY/*) and *ate* (*/EH/ /T/*). Because each syllable can only have one vowel sound, the system can assume that the vowel cluster "ea" in beaver maps to the phoneme */IY/*, whereas the "ea" in create maps to */IY/* and */EH/*.

2.1.2 Syllable Alignment

Unfortunately, because each of the syllabified data sets was developed independently using different sets of rules, the written and phonetic syllable boundaries are not always perfectly aligned in multi-syllable words. To determine how accurately the written and phonetic syllables initially aligned, the written syllables were manually compared to their matched phonemes for a set of words. Each of the words to test was chosen from an established phonics curriculum from the classroom of one of our research group's partner adult literacy organizations. Five random words were selected from each of the lessons in this curriculum that contained words of more than one syllable, resulting in 319 unique words and 670 individual syllables. Each syllable was evaluated independently as being

either correctly or incorrectly aligned, where an alignment was determined to be correct if and only if the written form of the syllable, when pronounced as written, was identical to the phonetic representation of that same syllable. An example of correct and incorrect alignments is shown in Figure 2-2.

mustang	mus → */M/ /AH/*	tang → */S/ /T/ /AE/ /NG/*
mundane	mun → */M/ /AH/ /N/*	dane → */D/ /EY/ /N/*

Figure 2-2. An example of a word with incorrect syllable alignments (top) and correct syllable alignments (bottom).

Each syllable was categorized according to its type, to better inform the evaluation and determine which specific characteristics result in the most misalignments. In English, there are six distinct forms that a syllable can take, each of which is described below:

1. A *closed* syllable ends with a consonant and has a short vowel sound (e.g. *cap*)
2. An *open* syllable ends with a vowel and has a long vowel sound (e.g. e-*cho*)
3. A *vowel-consonant-e* syllable ends with a silent ‘e’ that makes the vowel long (e.g. *bake*)
4. A *double vowel* syllable contains a cluster of vowels that make a single sound (e.g. *bound*)
5. A *consonant -le* syllable ends with “le” (e.g. puz-*zle*)
6. An *r-controlled* syllable ends with a vowel followed by an r (e.g. *her*)

A large portion of these misalignments were observationally determined to be the result of different consonant-vowel division rules between the two data sets, creating an off-by-one scenario between the coda of a closed syllable and the onset of the next. Closed syllables in the dictionary data set are consistently written with the consonant

sound as the coda, while the phonetic representation typically assigns this sound to the onset of the following syllable: for example, the word *pesto* is hyphenated as *pes-to* but phonetically split as */P/ /EH/ - /S/ /T/ /OW/*, creating an improper mapping wherein the *s* in the first syllable would be assumed to be silent and the *t* in the second syllable would map to *S T*. Doubled consonants are also consistently split between two concurrent written syllables (e.g. *pil-low*), while phonetically they represent a single unit (*/P/ /IH/ - /L/ /OW/*). Algorithm 2 describes in detail the adjustments made to the syllable onset and coda positions to achieve more accurate phonetic alignment.

Algorithm 2 Algorithm to adjust boundaries in written syllables to align with phonetic syllables

SB: list of all legal starting consonant blends
w: list of syllable/phoneme tuples (s, p)
phonemes: dictionary of phonemes in each word

```
for each (s, p) in w:
    if sₙ and sₙ₊₁ are in the dictionary:
        // if the word is a compound word, assume the given syllable
        // boundaries are correct
        if pₙ == phonemes[sₙ] and pₙ₊₁ == phonemes[sₙ]:
            (skip)

    if sₙ ends with a consonant:
        // consonant triplets get moved to the start of a syllable
        // (e.g. stretcher: stre-tcher / S T R EH - CH ER)
        if sₙ ends with a consonants triplet, move them to start of sₙ₊₁

        // if this syllable ends with the same consonant that the next
        // syllable starts with, combine them into the start of the next
        // (UNLESS "rr"/ER, then combine at the end of the previous)
        if sₙ[-1] == sₙ₊₁[0]:
            if sₙ[-1] == "r" and pₙ ends with /ER/:
                move sₙ₊₁[0] to end of sₙ
            else:
                move sₙ[-1] to beginning of sₙ₊₁

    else:
        // if the combined consonants are a legal starting blend, move
        // them to the second syllable
        if sₙ[-1] + sₙ₊₁[0] in SB:
            move sₙ[0] to beginning of sₙ₊₁
```

Table 2-1 shows the percentage of correct alignments for the same set of words after applying these adjustments. The application of these conditional onset-coda shift rules achieved an accurate alignment for over 96% of the syllables examined.

Table 2-1. The percentage of correctly-aligned syllables for each syllable type before and after adjustments

Syllable Type	**Size**	**Before**	**After**
Closed	*N = 358*	59.8%	95.0%
Open	*N = 101*	77.2%	99.0%
Vowel-consonant-e	*N = 32*	93.8%	100.0%
Double vowel	*N = 93*	76.6%	97.9%
Consonant -le	*N = 20*	70.0%	100.0%
R-controlled	*N = 66*	75.8%	98.5%
Total	***N = 670***	**68.3%**	**96.6%**

2.1.3 Evaluation

Using these more accurate syllable alignments, five words, both single and multi-syllable, were randomly selected from each lesson of the phonics curriculum and split along their corrected syllable boundaries, and each syllable was subsequently split into its onset, vowel, and coda. Each cluster was then examined individually, and every cluster was identified whose letters and phonemes were correctly linked.

Table 2-2. The percentage of correctly-mapped letter clusters by letter pattern

Pattern	**Size**	**Example**	**Correct**
C	*N = 660*	**h**-i-m	97.58%
CC	*N = 228*	d-o-**ck**	96.05%
CCC	*N = 13*	m-a-**tch**	86.67%
CV	*N = 47*	b-i-**te**	100%
V	*N = 464*	d-**o**-g	100%
VC	*N = 77*	b-**ow**-l	100%
VV	*N = 100*	s-**ea**-l	100%
Total	***N = 1592***		**98.30%**

Each cluster was categorized according to its letter pattern, generalizing each to a string of consonants (C) and vowels (V). The results for each of these patterns, and an example of each, are displayed in Table 2-2.

As can be seen from these results, this mapping algorithm, when coupled with the syllable alignment rules, is extremely effective at correctly identifying the relationship between the individual letters that make up a word and the phonemes that they produce for words that a typical literacy learner would be exposed to in a standard curriculum.

2.2 Grouping Words by Shared Features

When words are presented together in meaningful groupings, skills can be isolated in a very targeted way. Finding words that share a common feature can better challenge a student's mastery of the skills being targeted. For example, if a student is learning how to segment words into onset-rime pairs, it would be beneficial to be able to find words that all share the same onset, to provide a variety of examples of how this sound manifests in different words. In the same vein, there is also great benefit in being able to identify words that *differ* by a common feature. For example, a student who struggles with the differentiation between the /B/ and /D/ phonemes should be presented with interventions that specifically target this sound differentiation; while simply practicing words that contain those two sounds would be helpful, it would be significantly more effective to be able to find words that are identical *except for* this sound, to more pointedly target the weakness being addressed.

Grouping words by their shared characteristics can also be used for assessment purposes. Presenting a student with a prompt and providing three or four possible choices as an answer is a common method of assessing a student's understanding of a topic. To

make such questions most effective, the incorrect choices, or *distractors*, must be sufficiently close to the correct answer while still being identifiably incorrect to a student with full proficiency. A student who demonstrates mastery of an item should be able to select the correct answer from amongst the given distractors, while a student who is less proficient should find the distractors misleading and be prone to mistaking them for the correct answer. Selecting distractors that all share common features with the correct choice, or that differ from the correct choice in one specific way, is an effective method of ensuring that the distractors are sufficiently misleading so as not to allow a student to choose the correct answers by guessing and process of elimination.

The following sections describe three primary characteristics from which words can be compared and grouped: rhyming, orthographic similarity, and phonetic similarity.

2.2.1 Rhyming

A pair of words is considered a perfect rhyme if the stressed vowel sound and all other sounds following it are identical between both words. To find rhyming candidate words for a given target word, the system simply extracts the primary stressed vowel from the target word's phonemic breakdown and all the phonemes that come after it; then, to find words that rhyme with the original word, the system simply locates all other candidate words that end with that same ordered list of phonemes.

Several studies have found that adults had more difficulty detecting rhyming words that were not spelled similarly [37, 71]. Thus, as an increased challenge to the student, the system can also compare a word's phonetic makeup to its spelling: for example, the words *bead* and *head* share the same orthographic structure but do not rhyme, whereas the words *take* and *break* do rhyme despite their written dissimilarity. Orthographically-

parallel words are identified by examining letters including and after the last vocalized vowel cluster in each word; if the two strings are equal, the words are considered to have parallel spellings. The system can then compare the rhyming status of each pair to their orthography to find parallel or opposing pairs (see Figure 2-3).

(a)	*sing*	s - **ing**	/S/ – **/IH/ /NG/**
	ring	r - **ing**	/R/ – **/IH/ /NG/**
(b)	*bed*	b - **ed**	/B/ – **/EH/ /D/**
	head	h - **ead**	/H/ – **/EH/ /D/**
(c)	*pear*	p - **ear**	/P/ – **/EH/ /R/**
	hear	h - **ear**	/H/ – **/IH/ /R/**

Figure 2-3. (a) Words that are orthographically and phonetically parallel. (b) Words that are phonetically parallel but not orthographically. (c) Words that are orthographically parallel but not phonetically.

2.2.2 Orthographic Similarity

Orthographically-similar word pairs can be created by selecting words that differ from the target word by a single letter or letter cluster. Distractors of this format target a student's ability to differentiate between words that share similar written patterns.

At their most basic, orthographic groupings simply find words that share common letter patterns: for example, the words *bear* and *fear*, which are identically written except for their first letter. However, substituting letters according to their visual similarity can more accurately test a student's alphabetic knowledge and visual awareness: for example, the letters *b* and *d* are mirror images of one another, making *bear* and *dear* a more challenging pairing. Meanwhile, substituting letters according to their phonetic similarity tests a student's aural sensitivity, challenging their ability to distinguish between letters

with the same manner of articulation (e.g. *bear* and *tear*). Table 2-3 shows consonants and vowels categorized according to their manner of articulation.

Table 2-3. Letters in the English alphabet, grouped by the manner of articulation of their most common phonetic representation

Manner of Articulation	Letters
Vowels	a, e, i, o, u
Stops	b, d, g, k, p, t
Fricatives	f, h, s, v, z
Affricates	g, j
Nasals	m, n
Liquids	l, r
Semivowels	w, y

Orthographic groupings can also be used to identify words that share a common letter or letter cluster that is mapped to a different phoneme cluster. These types of items are extremely useful for targeting a student's understanding of how letters behave differently in different words, as is the case with long and short vowels.

Figure 2-4. An example of words that all share the same letter, but which exemplify a mix of different sounds for that letter (i.e. short and long vowel).

To find this class of candidate word for a given target word, the system simply finds all candidate words that share the same letter cluster at the same index as the target word. These words can then be classified according to which do and do not map to the same phoneme cluster at that same index. An example can be seen in Figure 2-4.

2.2.3 Phonetic Similarity

Phonetically-similar words can be chosen by selecting candidate words that have a phonemic breakdown that is partially the same as that of the target word. As distractors, these word groups target a student's ability to differentiate between words that sound similar when spoken aloud.

Words that differ from the target word by a single phoneme can be found by finding any word whose ordered list of phonemes is identical to that of the target word at all but one index. Words can be chosen that each differ from the target word at the same phoneme index to test a student's aural sensitivity to a specific phonetic sound, or they can differ by any one sound regardless of its position. Candidate words that share a phoneme with the target word, on the other hand, need only to have the same phoneme at the same index as the target word. Examples of these cases can be seen in Figure 2-5.

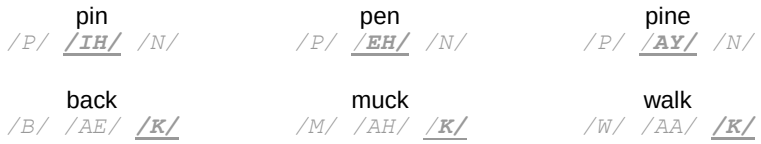


Figure 2-5. An example of words that differ from the target word by a specific phoneme (top), and words that share a specific phoneme with the target word (bottom).

Note that for phonetically similar words, the spelling of the resulting words is irrelevant. Some eligible words can differ dramatically in spelling (for example, *cough* and *calf*) while other words with very similar spellings (such as *whole* and *while*) may be ineligible due to their phonemic breakdowns.

Just as with orthographic groupings, phonetic groupings can also be used to identify words that share a common phoneme, but which map to different letter clusters. To find

this class of candidate word for a given target word, the system simply finds all candidate words that share the same phoneme cluster at the same index as the target word. These words can then be classified according to which do and do not map to the same letter cluster at that same index. An example can be seen in Figure 2-6.

Figure 2-6. Words that have a different orthographic representation for the same phoneme

2.3 Generating Word Misspellings

A student with a well-trained ear may be able to hear a pronunciation and identify it as a known word, but can struggle with the task of properly *encoding* that word: i.e., translating it into its correct written form. There are many different potential points of failure in the process of encoding a word, each of which ultimately results in a misspelling. A study from 1940 examined the spelling errors made by "poor" spellers, identifying the different types of errors that were most commonly made and categorizing them accordingly [72]. Derived from these findings, each of the algorithms described in this section is designed to test a different point of failure in the encoding process.

The generation process considers two distinct cases. The first case results in misspellings that, when sounded out phonetically, are not immediately obvious as being incorrect: for example, the word "sound" misspelled as *sownd*. These types of misspellings are useful for testing the reader's knowledge of the proper spellings of words, tapping into their memory and intuition of the English language. The second case results in misspellings that are identifiably incorrect when sounded out phonetically: for

example, the word “fish” (*/F/ /IH/ /SH/*) misspelled as *fisk* (*/F/ /IH/ /S/ /K/*). These types of misspellings target a student’s ability to decode a nonsense word and differentiate between its phonetic characteristics and the target pronunciation. [37] found that adult learners are more likely to make these types of non-phonetic spelling errors due to their heavier reliance on orthographic rather than phonetic cues.

Each of the rules for generating a type of misspelling is described in the following sections. Table 2-4 shows an example of the different misspellings that can be generated by the system.

Table 2-4. Examples of the types of misspellings generated for a variety of different words

	Drop letter	**Drop silent e**	**Transpose letters**	**Substitute phoneme**
speed	*sped*	--	--	*spead*
defeat	*defet*	--	*defaet*	*defiet*
corpse	*copse*	*corps*	*coprse*	*courpse*
repair	*repir*	--	*repiar*	*repear*

2.3.1 Omitting a Letter

Misspellings resulting from the removal of a letter are largely trivial to generate. When dropping an existing letter from a word’s spelling, any letter that is contained within a consonant or vowel cluster with more than one letter is treated as a compatible candidate (e.g. crack → *cack*, flies → *fles*). Constraining the dropped letters to come from multi-letter clusters avoids syntactically impossible misspellings (e.g. divine → *diine*, hold → *hld*).

The one exception to this rule concerns the removal of a silent “e” from the end of a word. A silent “e” is a non-vocalized vowel that lengthens the sound of the vocalized

vowel before it. Removing a silent “e” from the end of a word is valuable for testing the reader’s understanding of long and short vowels. However, not all words that end in “e” are good candidates for this removal: for example, *recip* would not be an intelligent misspelling for “recipe”, as this results in the removal of an entire syllable.

To determine whether an “e” at the end of a word is silent or vocalized, the system examines the last phoneme cluster of that word. It begins by finding all phoneme clusters that commonly map to the letter “e” using the calculated mapping frequencies. If the word ends with “e” and the final phoneme is not one of these sounds, the “e” is assumed to not be vocalized and thus can be removed from the word. For example, the word “scale” ends with the `/L/` phoneme, making *scal* an appropriate misspelling; on the other hand, the word “the” ends in `/IY/`, eliminating *th* as a misspelling.

2.3.2 Transposing/Reversing Letters

A common type of spelling mistake involves transposing or reversing the order of letters within a word. The system generates these types of misspellings by identifying every pair of consecutive letters and reversing their order (for example, “heart” becomes *haert*, and “helmet” becomes *hemlet*). The first two letters of a word are never reversed, nor are consonant digraphs which function as a single sound (e.g. th, ck, ng).

2.3.3 Substituting Letters

The system can generate phonetic misspellings for both consonant and vowel clusters within a word by utilizing the letter-phoneme mapping to find letter clusters that produce

the same phoneme(s) as the cluster being replaced. Figure 2-7 shows examples of the types of misspellings that can be generated using this method.

ch**ew** (/UW/) → *ch**u**, ch**oo***
ma**ke** (/K/) → *ma**ck**, ma**k**, ma**c**, ma**ch***

Figure 2-7. Examples of phonetic misspellings

First, the system determines the frequency distribution of every unique phoneme cluster in the database and every letter cluster it can map to. Any mapping that occurs less than 5% of the time is assumed to be either the result of an error in the mapping algorithm or an uncommon language case and is discarded. The result is a set of all unique letter clusters that have been regularly observed to map to each phoneme cluster.

To create a phonetic misspelling for a target word, the system iterates through each letter cluster within the word and extracts the phoneme cluster to which it maps. It then randomly selects a different letter cluster that maps to that same phoneme cluster (weighted according to its probability of producing said mapping), and substitutes that letter cluster in the target word. (See Appendix B for an example of the top three letter clusters that are mapped to each unique vowel phoneme in the ARPAbet phone set and their probabilities.)

2.4 Evaluation

The quality of the generated materials is evaluated using two metrics: correctness and coverage. A description of each of these metrics and their results can be found in the following sections.

2.4.1 Correctness

Incorrect distractors can be a severe detriment to the student; if the system claims that two words share a relationship that is not actually true, it reinforces incorrect information and interferes with the learning process. This paper asserts that a *correct* distractor is one that is easily identifiable by a literate reader. If a reader with well-developed literacy can consistently distinguish the target words from their distractors, the generated items are viable examples of tasks that proficient readers are able to accomplish, and thus they are appropriate tasks for teaching developing readers. This also ensures that the algorithms do not introduce false positive or false negative distractors.

2.4.1.1 Method

Simple surveys were employed to test the correctness of each of the distractor generation algorithms. Each survey item asked the participant to distinguish between two similar words, one being the target word and the other a generated distractor for that word. The survey was composed of four sections: rhyming words, orthographic distractors, phonetic distractors, and misspellings.

Respondents consisted of ten native English-speaking college graduates, with no known reading, learning, or hearing disabilities. Respondents were randomly paired into five groups, where each group was given the same random subset of questions to answer. Respondents' answers were tallied and compared against one another to determine the validity of the generated items. Items that were answered correctly by both of their respective respondents were determined to be *valid*, while *invalid* items were those for which both responses were incorrect. Questions that were answered incorrectly by one respondent but not the other were marked as *questionable*.

The rhyming pairs survey was structured as a series of yes-or-no questions, each asking: “Do these two words rhyme?”. The survey consisted of 150 word pairs in total, where 50 pairs rhymed and had parallel spellings (e.g. *hand* and *sand*), 50 pairs rhymed but had different spellings (e.g. *bowl* and *coal*), and 50 pairs had parallel spelling but did not rhyme (e.g. *pour* and *sour*). Participants were given 30 questions to answer in total, 10 of each type, in random order.

The surveys designed to test the orthographic and phonetic distractor generators presented pairs of words that differed from one another by a single letter cluster or a single phoneme, respectively. Respondents were asked to listen to the audio pronunciation of a word and to choose which of the two choices matched what they heard. 100 questions were generated for each type, and respondents were asked to answer 20 questions from each, for a total of 40 questions per respondent.

The survey for testing the misspelling generation algorithms presented participants with a single word and a misspelling of that word; respondents were asked to listen to the audio pronunciation of the word and to choose which of the two choices represented its correct spelling. Each misspelling was generated using one of the three different rules: transposing letters, omitting a letter, or substituting a phonetic cluster. 50 questions were generated for each rule, totaling 150 questions, and each respondent was given 10 of each for a total of 30 questions.

2.4.1.2 Results

Table 2-5 displays the total correct responses for items within each category, while Table 2-6 displays the distribution of valid, invalid, and questionable items generated by each of the four generation algorithms.

Table 2-5. The proportion of correct responses for each question type by all respondents

	Rhyming	**Orthographic**	**Phonetic**	**Misspellings**
Total Responses	300	200	200	300
Correct Responses	290	198	199	298
Correct Responses (%)	96.7%	99.0%	99.5%	99.3%

More than 97% of all items generated were answered correctly by all respondents. Only 2% of the generated items were considered questionable, and less than 1% were deemed invalid. These results strongly suggest that the generation algorithms are capable of producing items that can be correctly answered by literate readers, making them reliable benchmarks for comparison for low-literate learners.

Table 2-6. The percentage of each type of question deemed valid, invalid, and questionable

Category	**Total**	**Valid**		**Invalid**		**Questionable**	
Rhyming	150	141	94.0%	1	0.7%	8	5.3%
Orthographic	100	99	99.0%	1	1.0%	0	0%
Phonetic	100	98	99.0%	0	0%	1	1.0%
Misspellings	150	148	98.7%	0	0%	2	1.3%

The rhyming items section received the largest number of incorrect responses, with 141 of the 150 rhyming items being answered correctly by all respondents. Of the nine items that received an incorrect response, only one was determined to be invalid: *catch* and *snatch* were said to not rhyme by both participants, despite the CMUdict indicating that their phonemic breakdown is identical. It is unknown if this outcome was due to user error or perhaps influenced by a regional dialect. Of the questionable items, four were rhyming words that were spelled differently (*swan/drawn, sly/nigh, known/sewn,*

squat/dot), and four were non-rhyming words with parallel spelling (*raft/waft, mad/wad, chap/swap, glad/wad*).

The outputs from both the orthographic and phonetic generators proved to be almost completely correct. In the orthographic section, only one pair was found to be invalid: *expensive* was chosen as the written form of the audio for *expansive* by both respondents. As the phonetic representations of the "e" and "a" in these words sound very similar, this is not a surprising error. In the phonetic section, no questions were found to be invalid, and only one pairing was deemed questionable (*molt/mold*).

Of the misspelling items generated, only two were deemed to be questionable. One respondent identified the correct spelling of *ointment* as *ointmint*, a phonetic substitution producing a very subtle difference between the correct and incorrect spellings. The second respondent selected the misspelling *hadncuff* instead of *handcuff*, an error that could have been due to an accidental selection or an oversight. None of the misspelling items were found to be invalid.

2.4.2 Coverage

The previous section proved that the generation algorithms can create valid items that literate readers can reliably answer. However, in order for the generation system to serve as an effective tool for alphabetics instruction, it must be able to achieve sufficient coverage of a given curriculum and the skills contained within. It is therefore necessary to examine how much material the system can generate, how thoroughly each of the generators can cover the full range of skills within a curriculum, and at what difficulty levels each of the generators is most effective.

2.4.2.1 Method

To evaluate the generation algorithms for coverage, the output is tested against a popular early reading curriculum, the Wilson Reading System®. The Wilson Reading System is a 12-step remedial reading and writing intervention, where each step is divided into a series of sub-steps that introduce new phonological and orthographic skills of increasing complexity. The Wilson Reading System is advertised as an appropriate intervention for students in grades 2-12, as well as for adults with word-reading deficits or dyslexia, and has been used in many adult literacy programs since the late 1980s.

Two different metrics are utilized to evaluate each of the generation algorithms for coverage. First, it is necessary to examine how many of the words in the entire Wilson program are compatible target words for each generation algorithm. A word is considered to be a compatible target word if it is capable of generating at least one distractor of the requested format: for example, the word *orange* is not a compatible target word for the Rhyming Pairs algorithms, as it cannot produce any matching distractors. Second, it is necessary to examine how many unique distractors each of these target words can create on average: for example, when looking for words that differ by a single phoneme, the target word *pin* can produce many distractors by itself, such as *pen*, *pan*, *pain*, and *pine*. By examining how many words in the system can create distractors, and how many distractors these words can create, the applicability and thoroughness of each generation algorithm can be estimated for an average curriculum.

However, it is not enough to simply find all words in the Wilson curriculum that meet the distractor criteria for a given generator. A well-designed curriculum follows a carefully-crafted progression of difficulty, slowly introducing more difficult concepts to

the student only after they have sufficiently mastered the more foundational materials. A student who has just learned the consonant blend “br”, for example, should obviously be introduced to words like “bran” and “brim” long before they are exposed to “braggadocious”. The output of the generators must therefore be able to be constrained in order to filter out words that are beyond the appropriate difficulty level for any particular skill being targeted.

Unfortunately, there is no objective set of rules that specifies where different letter combinations and phonemes lie on a difficulty spectrum. There are numerous beginning reading systems on the market designed to introduce new readers to the alphabetic principle and teach them letter-sound relationships, and each system uses different approaches to this task, introducing letter combinations, phonemes, and syllable forms at different phases of the learning process. Because the generation system is designed to be program-agnostic and to accommodate any individual classroom’s preferred curriculum, the system cannot rely on hard-coded rules to specify which words and concepts are “easier” or “harder”. Instead, the system requires an instructor to specify their desired skill progression explicitly.

The system defines a *skill* as a unique combination of two distinct parameters: a letter-sound relationship, and a maximum number of allowed syllables. Relying on letter-sound combinations ensures that students are not being overwhelmed with words that are too phonetically or orthographically complex for their current level. For example, the words *thin* and *then*, with nearly identical spellings, might be introduced at different levels of the curriculum due to their different pronunciations of *th* (*/TH/* vs. */DH/*). Similarly, *pay* should not be paired with *eight* despite their shared vowel phoneme

(*/EY/*), due to their obvious difference in orthographic complexity. The results are also constrained based on the number of syllables allowed, ensuring that *banana* is not introduced in the same set as *ban*.

From these parameters, the system can automatically compile a set of every word in the database that targets this particular phonetic skill at the appropriate difficulty level, without introducing any words that contain skills that are specified to be beyond the level of a student learning this skill. This is accomplished using the method described in the following algorithm:

Algorithm 3 Algorithm to select words at each level from a specified skill progression.

Input:
P ← ordered list of phoneme skills

Definitions:
CP: set of all candidate phoneme skills
CW: set of all candidate words
TW: set of all target words

```
CP[0] = []
TW[0] = []
CW[0] = []

for i in 0...length(P):
   p = P[i]
   CP[i] = P[0] + P[1] + … + p
   TW[i] = all words containing skill p and only skills in CP[i]
   CW[i] = CW[i-1] + TW[i]
```

The result of this process is two ordered lists of word sets. Each candidate word set contains every possible unique English dictionary word made up of some permutation of all encountered phoneme skills up to that point, while each target word set contains all of these words that contain the current skill being targeted. While skills that have been previously encountered will always propagate forward to be included in later target sets,

no set of target or candidate words will ever incorporate phoneme skills that have not yet been encountered by the student.

The diagram in Figure 2-8 gives an example of a set of phoneme skills and the target and candidate word sets that result from the generation algorithm.

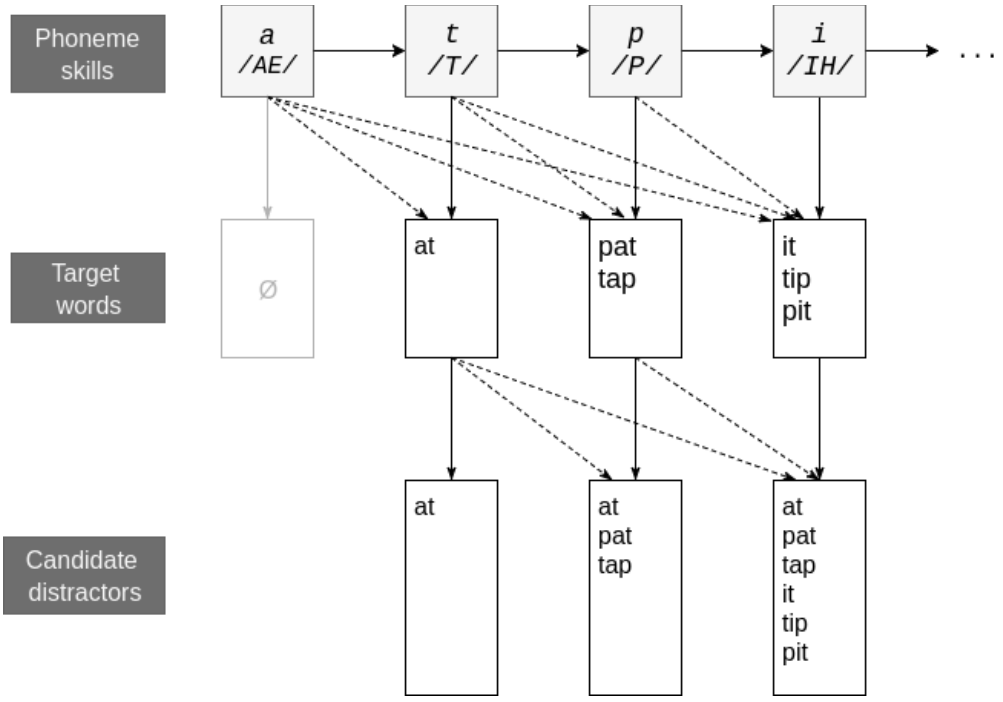


Figure 2-8. Diagram of the process of finding suitable target words and candidate distractors for each ordered phoneme skill

2.4.2.2 Results

Using the full set of words from all steps in the Wilson Reading System, coverage is evaluated by examining how many words are compatible target words for each generation algorithm, and how many unique distractors each of these words can create on average. These numbers are then compared against the results of the same algorithms

constrained to follow the specific skill progression of the Wilson Reading System, examining how many items are filtered out by the constraining process. For example, the unconstrained algorithm for finding words that differ by a single phoneme could pair the target word *got* (`G AA T`) with *bought* (`B AA T`), but *bought* would not be an acceptable distractor in the constrained curriculum due to its more complicated orthographic structure. (Note that once *bought* is encountered as a target word, the word *got* would be an acceptable distractor, allowing for the same concept to still be covered by the constrained curriculum, but simply delayed to a more appropriate time.)

Table 2-7. The percentage of all words that are compatible target words, and the average number of distractors a single target word can generate, using both restrictive and loose constraints for distractor selection

Format	Unconstrained	Constrained	% Filtered
Word Similarity			
Compatible Target Words	90.0%	89.0%	1.1%
Avg Distractors per Target	40.6	24.6	40.0%
Cluster Similarity			
Compatible Target Words	98.7%	98.7%	0%
Avg Distractors per Target	1314.4	338.1	74.3%

Table 2-7 displays the percentage of compatible target words across the entire Wilson word set and the average number of distractors (of any format) that a single target word can produce. Both tables are broken up according to the general format of the distractors: the "word similarity" format refers to any generation algorithm that pairs words with largely similar features (e.g. *pin* and *pen*, which differ by a single phoneme and a single letter), while "cluster similarity" refers to generation algorithms that target only a single cluster within each word pair (e.g. *slim* and *rice*, whose only similarity is the shared use

of the letter *i*). Because cluster similarity items have much fewer restrictions in finding a matching pair of words, these item types generate substantially more distractors for a single word, necessitating the division between these items and the more restrictive "word similarity" generators.

As expected, the cluster similarity items generated a much larger number of distractors for a single target word, and more than 98% of all words in the curriculum were capable of producing at least one distractor that shares some sort of cluster similarity. In comparison, roughly 90% of all words in the Wilson curriculum were found to be compatible target words for word similarity items. Few to no words were filtered out from either format during the constraining process, illustrating that even when the system limits the potential pool of distractors for a given target word, nearly every word in the system is still capable of being assessed to some degree through the distractor generation process. As the primary goal of the generation system is coverage of all input materials, this is a very positive and promising outcome.

2.4.2.2.1 Rhyming Pairs

There are multiple ways that the rhyming pairs generator can be applied to create distractors: the system can select pairs of words that rhyme and have parallel spelling, words that rhyme while having different spelling patterns, and words that have parallel spellings but do not rhyme. Each of these rules tests understanding of phonetic and orthographic patterns in a different unique way, and each requires a different set of criteria for generating appropriate items.

Table 2-8 and Table 2-9 outline the percentage of words that are compatible target words for each of these different rhyming formats, and the average number of distractors that each compatible target word can generate for each format, respectively.

Table 2-8. The percentage of all words that are compatible target words for each type of rhyme item

Format	Unconstrained	Constrained	% Filtered
Rhyme: Same Spelling	**56.2%**	**39.9%**	**63.9%**
Single syllable	78.0%	73.8%	52.7%
Multi-syllable	37.8%	9.5%	90.1%
Rhyme: Different Spelling	**39.4%**	**14.2%**	**30.7%**
Single syllable	60.2%	28.5%	5.4%
Multi-syllable	21.8%	2.2%	74.9%
No Rhyme: Same Spelling	**18.3%**	**9.6%**	**47.7%**
Single syllable	28.8%	19.4%	32.5%
Multi-syllable	9.4%	1.2%	86.9%

Table 2-9. The average number of distractors that can be generated from a given target word for each type of rhyme item

Format	Unconstrained	Constrained	% Filtered
Rhyme: Same Spelling	**8.6**	**6.5**	**71.8%**
Single syllable	9.4	7.1	66.6%
Multi-syllable	7.3	2.7	95.9%
Rhyme: Different Spelling	**5.2**	**4.1**	**47.8%**
Single syllable	6.1	4.3	28.5%
Multi-syllable	3.1	1.3	90.9%
No Rhyme: Same Spelling	**2.4**	**2.3**	**50.5%**
Single syllable	2.5	2.4	36.1%
Multi-syllable	2.2	1.0	93.7%

The results show that all three of the rhyming pair formats are substantially more effective for single-syllable words than for multi-syllable words. The most widely applicable format is for rhyming words that are spelled the same, covering nearly 40% of

all words in the constrained curriculum and over half of all words when unconstrained. The least applicable format targets words that are spelled the same but do not rhyme; in both the constrained and unconstrained item sets, only 10-20% of all words were capable of producing even a single distractor of the appropriate format.

Each compatible target word was also found to produce a fair number of distractors. Same-spelled rhyming target words were the most successful at producing distractors, with different-spelled rhyming target words also producing 4-5 distractors each on average. Same-spelled target words that do not rhyme were able to produce notably fewer distractors, likely due to the relatively uncommon existence of such words in English. As before, multi-syllable target words performed more poorly than single-syllable target words, especially after the curriculum was constrained; because words must have identical trailing phonemes to be considered to rhyme, it stands to reason that words made up of fewer syllables (and therefore fewer phonemes) would produce a greater number of rhyming pairs.

2.4.2.2.2 Orthographic and Phonetic Pairs

The generation algorithms that group words by their orthographic and phonetic similarity can be applied in several ways to create questions of different formats.

One application of the algorithms for finding phonetic and orthographic pair similarities is to find words that share a single common sound or letter cluster. There are three distinct ways that these word pairs can relate to one another. First, two words can both contain the same phoneme represented by the same cluster of letters: for example, *pole* and *home* both share the vowel phoneme */OW/* represented by the letter *o*. Second, words can share the same phoneme that is represented by a different cluster of letters:

pitch and *chop*, for example, both contain the phoneme */CH/*, produced by the letters *tch* in one word and *ch* in the other. Finally, words can share the same cluster of letters that do not produce the same phoneme: in the case of *rough* and *though*, the letters *gh* either produce the phoneme */F/* or are not vocalized at all.

Table 2-10 and Table 2-11 describe the percentage of compatible target words across these three formats ("same/same", "same/diff", and "diff/same", respectively), and the number of distractors each of these target words can produce on average. The results are further separated according to whether the shared cluster is a consonant or a vowel, to provide a more thorough picture of which types of clusters produce the most distractors.

Table 2-10. The percentage of all words that are compatible target words for each type of orthographic pairs and phonetic pairs item

	Consonants			**Vowels**		
Format	Unconstrained	Constrained	Filtered	Unconstrained	Constrained	Filtered
Same/Same	**84.8%**	**82.3%**	**3.0%**	**81.8%**	**78.2%**	**4.4%**
Single syllable	84.9%	84.9%	0%	84.1%	83.1%	1.1%
Multi-syllable	84.7%	78.0%	5.6%	79.9%	74.0%	7.4%
Same/Diff	**83.7%**	**64.8%**	**22.6%**	**84.3%**	**68.0%**	**19.4%**
Single syllable	83.3%	49.7%	40.3%	85.2%	64.2%	24.7%
Multi-syllable	84.1%	77.6%	7.7%	83.4%	71.1%	14.8%
Diff/Same	**75.9%**	**38.4%**	**49.4%**	**89.2%**	**76.5%**	**14.2%**
Single syllable	70.6%	21.2%	70.0%	85.0%	69.1%	18.8%
Multi-syllable	80.3%	53.0%	34.0%	92.8%	82.9%	10.7%

When unconstrained, roughly 80-90% of words were compatible as target words for each of the three formats. Once constrained, roughly 20% of these words became no longer valid for producing same-phoneme/different-letter pairings for both consonants and vowels. For different-phoneme/same-letter pairings, the constrained data set filtered out nearly 50% of the compatible target words for consonants, but only approximately

14% of vowel target words were invalidated. This is not surprising, as it is more common for vowels to produce different sounds (e.g. long and short vowels) and for these concepts to be introduced early in a curriculum, while different-sounding consonants tend to be exceptions that would not be introduced until a more advanced reading level. Additionally, these formats tend to perform better for multi-syllable words, due to the greater number of clusters available in these words to be matched against other words.

Table 2-11. The average number of distractors that can be generated from a given target word for each type of orthographic pairs and phonetic pairs item

	Consonants			**Vowels**		
Format	Unconstrained	Constrained	Filtered	Unconstrained	Constrained	Filtered
Same/Same	**267.3**	**88.6**	**67.9%**	**158.4**	**69.8**	**57.9%**
Single syllable	213.2	84.6	60.3%	207.7	104.2	50.4%
Multi-syllable	313.3	92.2	72.2%	114.3	36.9	70.1%
Same/Diff	**192.9**	**69.7**	**72.0%**	**102.3**	**35.9**	**71.7%**
Single syllable	122.3	73.0	64.3%	77.1	32.1	68.7%
Multi-syllable	252.2	67.9	75.2%	124.1	38.7	73.4%
Diff/Same	**55.6**	**24.5**	**77.7%**	**726.1**	**135.2**	**84.0%**
Single syllable	13.4	17.5	61.0%	229.8	100.7	64.4%
Multi-syllable	87.0	26.9	79.6%	1111.2	159.6	87.2%

Because these pairings do not require there to be any similarities between two words other than their shared cluster/sound, items of these format have the potential to generate a very large quantity of distractors for a single word, even when the results are constrained. The constraining process filters out approximately 60-85% of distractors for each format, but even after this, the average number of distractors for *each* target word ranges from roughly 24 to 135. Once again, multi-syllable words perform better than single-syllable words due to their greater number of clusters to be matched against. As every target word has three possible phoneme-letter clusters for every syllable it contains

(where each syllable is made up of onset-vowel-coda triplets), and each of these clusters can produce its own set of distractors, words with more syllables have a greater chance of producing more matching items.

Another application of the algorithms for finding phonetic and orthographic pair similarities is to find words that differ from one another by a single phoneme or letter cluster. These items are far more restrictive than the previous format, requiring the two words to be identical beyond their single different feature. As such, we expect significantly fewer items to result from pairings of this type.

Table 2-12 and Table 2-13 break down the percentage of compatible target words for words that differ by a single phoneme and words that differ by a single letter cluster, and the average number of distractors each of these words can produce, respectively. Results are further divided according to whether the differing feature is a consonant or a vowel.

Table 2-12. The percentage of all words that are compatible target words for word pairs that differ by one phoneme and one cluster

	Consonants			**Vowels**		
Format	Unconstrained	Constrained	Filtered	Unconstrained	Constrained	Filtered
One Phoneme	**58.1%**	**38.0%**	**34.7%**	**38.7%**	**26.7%**	**31.0%**
Single syllable	84.9%	80.3%	5.4%	70.4%	57.4%	18.5%
Multi-syllable	35.5%	2.1%	94.1%	11.8%	0.7%	93.7%
One Cluster	**58.7%**	**38.5%**	**34.3%**	**31.0%**	**22.9%**	**26.2%**
Single syllable	84.8%	80.8%	4.8%	59.0%	49.5%	16.1%
Multi-syllable	36.5%	2.8%	92.4%	7.3%	0.4%	94.6%

The results show that there is a substantial difference in how applicable these types of items are for single- and multi-syllable words: when the curriculum is constrained, less than 3% of all multi-syllable words were found to be compatible target words for both consonant and vowel features, while 50-80% of single-syllable words were compatible,

depending on the format. This is not unexpected, as the more syllables a word contains, the more restrictive the matching algorithm becomes: words with one syllable need only to match two out of their three letter and/or phoneme clusters, while two-syllable words must match on five out of six, three-syllable words on eight out of nine, etc.

Table 2-13. The average number of distractors that can be generated from a given target word for word pairs that differ by one phoneme and one cluster

	Consonants			**Vowels**		
Format	Unconstrained	Constrained	Filtered	Unconstrained	Constrained	Filtered
One Phoneme	**16.0**	**12.7**	**48.1%**	**3.5**	**2.3**	**56.1%**
Single syllable	22.7	13.1	45.6%	4.0	2.3	53.7%
Multi-syllable	2.4	1.5	96.4%	1.3	1	95.0%
One Cluster	**16.6**	**13.5**	**46.6%**	**2.3**	**1.8**	**56.1%**
Single syllable	23.9	13.9	44.3%	2.5	1.8	40.2%
Multi-syllable	2.2	1.3	95.4%	1.1	1	95.3%

Words differing by a single consonant phoneme or letter cluster were able to produce significantly more distractors for each target word than vowels, likely due to the greater number of consonants to choose from: there are only 15 vowel phonemes in the CMUdict compared to 24 consonant phonemes. Additionally, a consonant cluster can be matched in the onset *or* coda of every syllable, where vowels can be matched only in the middle of a syllable. Despite this, single-syllable target words were found to produce roughly 2 vowel-based distractors each, and closer to 13 consonant-based distractors each, making this still an effective generation method for producing assessments.

2.4.2.2.3 Misspellings

Unlike the previous sections, the algorithms for generating misspellings do not rely on pairing real words together based on shared features: target words are simply

transformed into different, non-real words. As such, there is no difference in the results between a constrained and unconstrained curriculum, because the distractors do not come from the curriculum itself.

Table 2-14 outlines the percentage of compatible target words and the average number of distractors that can be generated for each of the three misspelling rules: transposing letters, omitting letters, and substituting letters phonetically.

Table 2-14. The percentage of all words that are compatible target words and the average distractors generated by each target word for each misspelling rule

Misspelling Rule	Compatible Target Words	Distractors per Target (Avg)
Transposition	**71.6%**	**1.5**
Single syllable	58.1%	0.8
Multi-syllable	81.6%	2.0
Omission	**68.6%**	**1.6**
Single syllable	56.9%	1.0
Multi-syllable	77.4%	2.0
Phonetic	**84.4%**	**2.2**
Single syllable	68.6%	1.0
Multi-syllable	98.7%	3.1

Across the entire Wilson curriculum, roughly 70-85% of words were found to create at least one type of misspelling. On average, each rule was found to produce one misspelling for every single-syllable target word, and 2-3 misspellings for multi-syllable targets. Phonetic substitution misspellings were applicable to the largest number of words, and were also capable of producing the most distractors per target word.

2.5 Future Work

The work described in this chapter largely focused on the alphabetics generation system's potential as a tool for instructors to create high-coverage, high-quality learning materials for their students. The system was evaluated based on how thoroughly it incorporated the words and skills within a given curriculum, and on how few incorrect or misleading items it generated. However, all of this evaluation was conducted "within a vacuum": to truly test the system's efficacy as a content creation tool, it must be evaluated in the actual environment in which it is intended to be used.

In order for the tool to be able to generate materials to supplement classroom learning for students, it is imperative that the content creators (i.e. the instructors) be able to harness the tool easily and effectively to create the content they want to see. Future work will test the usability of the creation system from the perspective of an end user, testing whether real instructors given the tool are able to generate the content they desire. This will expose both potential issues in the way that the tool is presented to the end user, as well as possible discrepancies between the output expected by the instructor and the output returned by the generators. This information will help to inform the design of the instructor-facing component to eliminate obstacles that could stand in the way of the content generation itself.

It is also important to assess the quality of the generated content from a more subjective standpoint. Although this thesis has examined the validity of the generated items from the perspective of how little false or misleading information was incorporated into the final results, future work should seek to evaluate these items according to how they compare to hand-created materials commonly used in other programs. Ideally, the

output from the generation algorithms should be indistinguishable in quality from materials created by knowledgeable experts; the system is designed to assist instructors in creating materials faster and more thoroughly than they can do by hand, but is not intended to produce results that are *better* than human output. Future work should explore how both instructors and students perceive the items generated by the algorithms in comparison to hand-crafted items, to ensure that there is no discrepancy in quality or reception by real users.

Finally, future work should seek to determine the effectiveness of the generated items as learning materials. It is important to ensure that the materials output by the system are as capable of imparting the intended skills to the user as materials currently being used in existing literacy programs. Testing this will involve conducting a long-term learning study over several weeks or months, wherein students are given regular practice with the generated materials and their proficiency is measured before and after, using standard assessment tools. It is the hope that the generated materials would prove to be at least as successful as hand-created items in improving assessment scores over time, confirming their effectiveness as a learning tool.

2.6 Summary

This chapter outlined a system for automatically generating reading exercises from an existing set of materials to target phonemic awareness and word knowledge. The system can generate items to test a reader's understanding of rhyming sounds and letter patterns, the ability to distinguish between individual phonemes and letter groupings, and how to properly spell words and identify incorrect spellings. The results of human evaluations show that the generated results are almost entirely valid as learning materials, producing

few to no incorrect or ambiguously correct results that a literate reader would not be able to identify. Additionally, extensive analysis of the output of the generation algorithms when run on an established early-reading curriculum shows the system's ability to generate a wealth of learning materials for each of the different skills described above, even when the output is strictly constrained to only allow for level-appropriate distractors. As such, this chapter concludes with the assertion that it is feasible for a single suite of generation algorithms to achieve sufficient coverage of all alphabetics skills, and that the system developed here has the potential to be a very impactful and effective educational tool for beginning-level adult readers.

Chapter 3 - Generating Comprehension Items

The ability to go beyond the explicit meaning of a text and to make educated assumptions about the intended meaning not directly stated is a critical reading skill that poor readers consistently struggle with compared to skilled comprehenders. This process of inferring implied information requires the reader to identify the important words in the text, mentally activate the necessary facts about those words, and build relationships about these facts through reason [84]. [85] hypothesized that skilled readers instinctively monitor their own comprehension, allowing them to identify when necessary information is missing and attempt to fill these gaps with inferences. Poor comprehenders, on the other hand, tend to approach reading as a task of decoding accuracy rather than deeper awareness. As a result, when readers fail to form an accurate mental representation of a text, they are unable to recognize blatant contradictions and inconsistencies [86], even when they are able to refer back to the text [87].

This chapter discusses the design of two question generation systems for creating practice materials at the comprehension level, both of which target a reader's ability to draw inferences in different ways. Section 3.1 begins with a brief overview of previous work in reading-based question generation systems. Section 3.2 describes a novel algorithm for introducing inconsistencies into an existing text which maintain local consistency while disrupting the logical meaning of the entire passage, and outlines the results of a human evaluation on the quality and reliability of the generated items. Section 3.3 describes an algorithm for generating items to challenge a reader's ability to understand the function of different discourse connectives, and evaluates the validity of

the resulting items. Finally, Section 3.4 concludes with a brief discussion of the contributions of this chapter and a summary of the findings.

3.1 Previous Work

Several previous studies have described systems for automatically creating exercises to target different aspects of language learning and vocabulary assessment. [73] describes a method of generating distractors for assessing an ESL reader's ability to distinguish semantic nuances between vocabulary words. [74] utilizes WordNet word relations and frequencies to generate distractors for vocabulary words from equally-challenging terms. [75] and [76] both explore methods of generating distractors of different classes designed to indicate specific deficiencies in phonetic or morphological vocabulary mastery. Others focus on generating exercises for quizzing or knowledge testing purposes. [77] explores generating gap-fill exercises from informative sentences in textbooks, while [78] locates suitable distractors for medical texts from domain-specific documents. Both methods choose distractors from other sentences in a constrained set of source texts rather than relying on external corpora.

A few studies have focused on more comprehension-specific exercises, generating distractors that are semantically similar to the target word. [79] proposes a method of generating semantically-similar distractors to the target word using context-sensitive lexical inference rules. The distractors generated using this method are contextually and semantically similar to the target word, but not in the context being used in the sentence. The RevUP system described in [80] utilizes a word vector model trained on the desired text domain to find semantically-similar words and verifies their similarity using WordNet synsets. [81] generates semantically-similar distractors using distributional data

obtained from the British National Corpus, and utilizes the Google *n*-grams corpus to determine each generated distractor's probability of occurring with its surrounding terms.

[84] describes the creation of a system called DQGen which generates cloze questions for testing different types of comprehension failure in children, including one method which creates "plausible" distractors that create contextually sensible sentences in isolation but do not fit in the context of the rest of the text. Their system also utilizes the Google *n*-grams corpus for finding semantically consistent distractors for these sentences. However, their attempt to generate distractors at the sentence level that are contextually inconsistent at the passage level returned limited results, as most target words were found to be easily distinguishable without needing previous sentences for context.

3.2 Identifying Locally-Consistent Inconsistencies

Similar to the type of comprehension failure assessment described by the DQGen system, the generation algorithm described in this section seeks to produce a form of comprehension monitoring exercise that introduces "locally-consistent inconsistencies" into the text, testing contextual understanding and the reader's ability to identify mismatches between the text they read and the mental model they build. This paper asserts that a locally-consistent inconsistency should make sense both grammatically and logically within its surrounding narrow context, but should not make sense within the broader context of the text. This type of exercise encourages engagement and focus while reading: because a well-formed item should not have obvious inconsistencies within a narrow reference frame, the reader must actively construct meaning and incorporate it into their mental model to identify the correct answer. Figure 3-1 gives an example of the type of item that the generator should produce: when looking at a narrow context, all the

word choices are logical selections for the blank, but when the meaning implied by the surrounding text is considered, only one choice is sensible.

(a) …*to stay* {**open** / **safe** / **quiet** / **down**} during…

(b) *Keep away from windows to stay* {open / **safe** / quiet / down} *during a hurricane.*

Figure 3-1. (a) In a narrow context, all four word choices are equally fitting; (b) In the full context, only the target word logically fits

A unique application of the Google Books *n*-grams corpus [83] is explored for generating reasonable locally-consistent distractors for a blanked word. Google *n*-grams is a massive corpus containing frequency counts for all unigrams through 5-grams that occur across all texts in the Google Books corpus.

The system begins by gathering every 2- through 5-gram in the original sentence that contains the target word. If the sentence contains multiple clauses, only the clause which contains the target word is considered. The system then employs a sliding window to gather all *n*-grams ($2 \leq n \leq 5$) within the clause of the form $\{w_1 \ldots w_{t-1}, [w_t], w_{t+1} \ldots w_n\}$, where the target word $[w_t]$ occupies each position $1 \leq t \leq n$. The entire Google corpus is then queried for *n*-grams matching each pattern $\{w_1 \ldots w_{t-1}, [w_t.pos], w_{t+1} \ldots w_n\}$ $(1 \leq t \leq n)$, where $w_t.pos$ represents the part of speech of the target word w_t (obtained using the Stanford Part-Of-Speech Tagger [94]). If the query returns no results, the system attempts to generalize the pattern further by replacing proper names and pronouns with their part of speech (see Figure 3-2).

Distractor queries follow a back-off model, using *n*-grams of size $n = \{5 \ldots 2\}$. For each n-sized pattern searched, the system identifies the intersection D of all words at index t (limiting the results to the top 100 for the sake of performance).

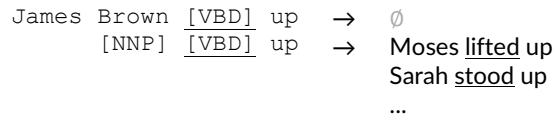


Figure 3-2. When n-gram queries return no results, specific terms are generalized to increase the likelihood of finding a match

None of the generated distractors should fit the blank as effectively as the target word, necessitating the removal of all words in D that are likely to make too much sense in context. Because synonyms can often be used interchangeably in the same sentence, all words are discarded that are direct synonyms of w_t (identified using synsets gathered from WordNet[3]). From the final set D', the system selects the three least-frequently occurring distractors in the Google corpus.

3.2.1 Identifying Contextually-Relevant Words

The previous section described the method of selecting distractors for a generated blank. However, not every word would make a useful question. The system needs to specifically prioritize words that are contextually relevant to the meaning of the passage; if the reader can infer the sentence's intended meaning with this word removed, then the task of replacing the word should be straightforward. The system considers a word to be contextually relevant if there are enough context clues in the surrounding text for the

[3] http://wordnet.princeton.edu

reader to understand the text's intended meaning even when the chosen word has been removed or replaced.

The system begins by considering every content word in a text to be a potential target word. Function words (articles, pronouns, conjunctions, etc.) are discarded from the pool due to their closed nature and frequent appearance across documents. However, unlike several other studies (e.g. [88, 89]), content words are not eliminated based on their global word frequencies: target words that successfully challenge comprehension of the surrounding context should implicitly test mastery of the more challenging words in the passage, regardless of the difficulty of the word itself, because a text cannot be fully comprehended unless the reader can parse and understand the vocabulary. However, *local* word frequencies are considered, and words whose stemmed form appears in the document multiple times are eliminated to ensure that readers cannot identify target words based solely on short-term memory recognition.

Because these exercises are designed to test comprehension rather than background knowledge, questions are not intended to "quiz" readers on facts, as is the case in many other studies (e.g. [90]). Therefore, classes of words that typically present factual information and could be easily exchanged for any other word of the same class, specifically named entities and numbers, are discarded as potential target words (as demonstrated in Figure 3-3).

In the fall of 2012, the New York City government began receiving unusual complaints.

By the time California became a state, it was already an important place for farming.

Figure 3-3. Examples of poor target words

After this filtering step, the remaining set of words serves as the pool of *potential targets*.

3.2.1.1 Contextual Scope

Individual sentences in a passage are rarely conceptually independent from one another. True understanding of a sentence's meaning often relies on information that has been gathered from previous sentences in the passage.

[91] found that traditional cloze-style comprehension questions are not good indicators of "intersentential comprehension", the ability to process and apply information across sentence boundaries. The ability to integrate previously-read information into one's mental model and carry this information through to later sections is a necessity for making inferences about a text and identifying inconsistencies, both of which are critical skills for comprehension. [92] verified this fact in studies that asked readers to resolve anomalies in a written text: they found that readers of all skill levels were equally able to identify and resolve textual inconsistencies when the resolving information was in an adjacent sentence, but when the information was further apart in the text, poor comprehenders performed significantly worse.

To attempt to address this issue, the system explores several different contextual "scopes" when attempting to find pairs of words with contextual links. Adjusting the scope of included information allows the method of selecting target words to incorporate potentially relevant or necessary context words that a reader has internalized from the sentences they have already read, challenging their intersentential comprehension and mental modeling. Potential context words for a given target word were chosen from only the target sentence, as well as from the target sentence and one or two sentences previous.

The pool of scope words for each sentence is filtered less rigorously than the pool of potential targets, as many of the word classes that make poor target words are poor choices specifically because they provide important facts that can be leveraged for context. Therefore, only function words are removed from the pool, leaving named entities, numbers, and frequently-occurring words as potential context words.

3.2.1.2 Word Co-occurrences

By definition, words that co-occur regularly are likely to have a contextual and/or semantic relationship to one another. The system therefore leverages co-occurrence likelihoods between words to select the potential blanks with the strongest relationship to their scope-specific context words.

Word co-occurrence likelihoods are represented using the word vector space model GloVe [93], trained on 42-billion tokens. The GloVe model formulates word vectors such that the dot product of any two word vectors $\widehat{w}_1 \cdot \widehat{w}_2$ represents the logarithm of the two words' probability of co-occurring together in a document.

The goal is to find the scope word for each potential blank with the highest likelihood of co-occurring with that blanked word. Using the GloVe model, for each potential blank $b \in B$, the system locates the closest scope word c in the set of all scope words S for that blank such that $(b, c) = arg\ min\ (\hat{b} \cdot \hat{c})$ ($\forall\ s \in S$ such that $s.stem \neq b.stem$). Each of these pairs is added to the pool of contextually-relevant words.

The system also uses these co-occurrence likelihoods to eliminate all potential distractors $d \in D$ such that $(\widehat{w}_t \cdot \hat{d}) < (\widehat{w}_t \cdot \hat{c})$ (where c is the closest scope word in the

pair (w_t, c)), because these words have a *higher* likelihood than the target word of co-occurring with their context words.

3.2.2 Evaluation

To evaluate the quality of the generated items, respondents were asked to answer fill-in-the-blank multiple-choice questions where the inconsistencies were presented as choices alongside the target word. 120 reading comprehension text passages were randomly selected from the corpus at ReadWorks.org[4], ranging in Lexile level from 100L to 1000L, and a single sentence was selected from each passage, presented to the participants as a multiple-choice question with the target word and three locally-consistent distractor words as choices.

The questionnaire was separated into two sections, both of which asked participants to answer the blanked multiple-choice questions. The first section presented each question at the phrase level (i.e. the blank surrounded by a small subset of the words in the full sentence). The words to include in these phrases were selected by hand to present the blank in a representational narrow context. The second section presented the full blanked sentence, surrounded by the context of the entire passage (or, in the case of particularly long passages, by relevant paragraphs from the full text). For both sections, participants were presented with four word choices for each blank, and were asked to select *all* of the words they believed logically fit the blank.

67 native English-speaking volunteers provided their feedback through an anonymous online questionnaire. Each participant was given a random subset of questions from each section to answer: 20 phrase-level questions, and 10 sentence-level questions.

[4] http://www.readworks.org/

Participants were not aware that the questions were generated automatically and were not informed of the research objectives or what we hoped to obtain from their answers, to avoid potential feedback bias.

3.2.3 Results

The content validity of the questions and the chosen distractors was determined by examining the proportion of words that fit each blank in a narrow context to words that fit the same blank in the broader context of the surrounding text. In an ideal question, the target word and all distractors should fit in the narrow context, and only the target word should fit given the full context. Thus, for target words, the aim is for 100% fit in both contexts; for distractors, 100% should fit in the narrow context and 0% in the full.

As can be seen in Table 3-1, the proportion of distractors deemed to fit the blanks in a narrow context increases substantially as *n* increases, while the proportion of target words chosen to fit is relatively unaffected. This pattern also holds true given the full context, although to a lesser extent.

On average, 58% of all distractors generated were deemed to fit in their given blanks in a narrow context, although this number is skewed by the poor performance of the bigram model. The 5-gram model was the best-performing for finding distractors that fit in the narrow context, achieving an average fit of approximately 74%. As *n* increases, more of the syntactic and semantic features of the phrase can be incorporated into the distractor selection, increasing the chances of the selected word making both grammatical and contextual sense with *all* of the words in the phrase.

Table 3-1. The percentage of distractors and target words chosen to fit each blank given the narrow context and the full passage

	Narrow		Full	
	Distractor	**Target**	**Distractor**	**Target**
n = 2	30.9%	93.1%	3.1%	98.0%
n = 3	57.7%	92.4%	7.0%	93.8%
n = 4	67.1%	89.9%	21.9%	95.5%
n = 5	74.1%	93.2%	13.2%	91.3%
All	**58.0%**	**92.1%**	**11.6%**	**94.6%**

Less than 12% of all distractors on average were deemed to fit the same blanks when given the full context, though the 4-gram model had the worst performance with nearly 22% fit. The bigram model performed best in the full context with approximately 3% fit; however, its poor performance in the narrow context suggests that these words are obviously incorrect and therefore not suitable distractors.

Table 3-2 compares the proportions of distractors fitting within each context across both *n*-gram model and scope (*s1* through *s3*). The same pattern of increasing fit with higher values of *n* can be observed within each scope. However, the scope does not appear to have a significant effect on the quality of the distractors generated.

Table 3-2. The percentage of distractors fitting each blank given the narrow and full context, for each scope

	Narrow			Full		
	s1	***s2***	***s3***	***s1***	***s2***	***s3***
n = 2	29.5%	31.9%	27.9%	3.2%	3.0%	3.4%
n = 3	53.7%	61.2%	61.0%	4.6%	10.2%	11.1%
n = 4	64.8%	66.7%	66.1%	25.3%	18.9%	21.5%
n = 5	75.7%	75.9%	75.0%	13.6%	13.9%	20.6%
All	**56.2%**	**59.4%**	**56.9%**	**10.4%**	**11.9%**	**13.5%**

Results suggest that larger *n*-grams are significantly more effective in creating sensible distractors that make sense within a narrow context, and that a large portion of

these distractors become no longer suitable once the full context of the passage has been introduced. This suggests that this method is a promising first step towards the generation of these types of comprehension-challenging exercises.

3.2.4 Limitations and Future Work

The proportion of words deemed to fit in the narrow contexts was lower than expected for both target words and distractors. It is possible that the concept of words "fitting" in a sentence fragment may not have been fully understood by some participants. For example, many respondents said that the word *went* was not a suitable fit for the phrase *Hidalgo _____ about this*. In this case, some participants may have struggled to identify the phrasal verb "to go about" as being grammatically correct because it clashed with the other choices (*heard, agreed, said*), where they might have chosen it to fit if it had been presented independently. A future study will explore a less subjective method of evaluating target words within a narrow context.

Perhaps the biggest weakness in the current method lies in filtering out fitting distractors. As indicated in the results section, approximately 12% of all the distractors generated using the algorithm were deemed to make as much sense in context as the target word. The majority of the distractors chosen to fit within their full contexts were observationally found to be "near-synonyms" of the target word (for example, the words *turned* and *flushed*, which are not obvious synonyms but are interchangeable given the context of the phrase *her face _____ red*.) While WordNet was employed to remove direct synonyms, a more robust synonym-filtering process seems necessary for future work, taking advantage of the already-utilized corpora to identify semantically-similar word pairs.

Further exploration should also be done into how sentential scope affects the target word selection process. The evaluation in this thesis examined the effectiveness of finding contextually-relevant words to the target word when considering context words from the same sentence and up to two sentences before the target, finding that the scope had no discernible effect on the validity of the items generated. Future work should seek to determine if further extending the scope to include even earlier sentences, or incorporating subsequent sentences into the scope, would have a more noticeable effect on item validity. It would also be beneficial to explore whether the target words selected are more, or less, suitable for targeting comprehension as the scope of available context is modified, necessitating the development of an evaluation method for determining the quality of the target words selected.

Additionally, it is important to explore whether the content of the passages used influences the validity of the output. Future work will seek to determine whether the reading level of the passage, or its format or content (e.g. newspaper article, literary story, poem), has any noteworthy impact on the output of the generators and the success of the algorithm.

Finally, alongside improvements to the question generation algorithm's performance, future work should also seek to prove the efficacy of these types of exercises in targeting the reading comprehension and inference-making skills of the intended user base. This process will involve further user evaluation, this time involving low-literate readers.

3.3 Applying Connectives

Cohesive devices are frequently used to aid in the integration of information between successive sentences and clauses, providing critical scaffolding to allow for the

production of inferences [50]. Discourse connectives are cohesive devices used to imply the relationship between otherwise disconnected parts of a text. Different types of connectives are used to imply different coherence relations: for example, words like "before" and "after" indicate a temporal connection between two events, while words like "because" and "so" indicate a causal relationship. The application of the wrong connective, or the misinterpretation of a connective's function in a text, can change the entire meaning behind what is read.

For example, the sentence *Sarah ate breakfast and she went to work* tells the reader very little about Sarah: all we know is that she performed two actions at some time, but the significance between these two events is unknown. However, the sentence *Sarah ate breakfast because she went to work* introduces several implications: as readers, we can infer that Sarah does not usually eat breakfast, and that Sarah likely works a physically demanding job requiring her to eat before she goes. While neither of these facts were explicitly stated, the use of the connective "because" to link the two clauses of this sentence provides us with key information to fold into our situation model and allow us to gain a better understanding of the meaning behind the text.

Research has shown that poor comprehenders consistently struggle with the proper application of discourse connectives [96, 97], particularly those that encompass temporal, causal, and adversative relations [98]. Being able to correctly interpret and apply connectives when reading is imperative for comprehension and the accurate maintenance of a reader's situation model. As such, we wish to generate exercises that target a reader's understanding of connectives and their functions.

To accomplish this, all the connectives in each text are identified by utilizing an end-to-end discourse parser developed by [99] which is modeled on the Penn Discourse Treebank, a large corpus of English texts annotated with their discourse structure and semantics [100]. The parser is designed to identify all implicit and explicit discourse connectives within a text, including their semantic classes and their relation type. Semantic classes are grouped into one of four categories: Temporal, Expansion, Contingency, and Comparison. Each of these categories encompasses a set of subtypes that describe their role in more detail: for example, a contingency connective can be one of *cause* (e.g. "Everything changed *when* the Fire Nation attacked.") or one of *condition* (e.g. "I won't be hungry *if* I eat this sandwich."). The semantic class of a connective and its subtypes dictate what purpose it serves within a sentence and the relationship that it implies between two connected clauses.

With this information, it is then possible to create questions to test both a reader's understanding of the meaning of connectives and their ability to infer the relationship between two pieces of text. The generation system begins by parsing a given passage and extracting all *explicit* discourse markers within. If the parser identifies that a discourse marker contains an explicit discourse connective, that connective can then be removed from the original text and turned into a multiple-choice question. Each of these connectives is compiled into a set of potential distractors, grouped according to their class and subtype. Once all candidate connectives have been removed from the text, distractors are then randomly selected from the pool of all possible connectives, where each distractor is required to have either a different class than the target word, or be of a different subtype of the same class, as connectives of the same subtype within the same

class are often largely interchangeable in meaning (e.g. temporal-synchronous connectives, as in the sentence: “He went to the store *when/while/as* I took a nap.”).

3.3.1 Evaluation

To evaluate the quality of the items generated by the connectives algorithm, ten native English-speaking college graduates with no known reading disabilities were asked to answer a series of randomly-generated multiple-choice fill-in-the-blank questions. 100 questions were generated from a set of reading comprehension passages obtained from the Marshall Adult Education project[5], ranging in difficulty from CASAS levels 200-235. Respondents were paired into five groups, with each group given the same 20 questions to answer, and their answers were tallied and compared against one another to determine how many of the generated distractors were valid, invalid, and questionable.

Each question took the form of a fill-in-the-blank multiple choice, where a connective was removed from the original text and presented as one of four possible choices, along with three other connectives of a different class and/or subtype. Respondents were asked to select *all* connectives that would logically fit in the given blank in each sentence, allowing for more than one answer per respondent per question.

3.3.2 Results

Table 3-3 shows the total number of responses received by all respondents, and the ratio of correct to incorrect responses. Note that, although there were 100 questions and each question was answered two times, a question could receive more than one answer,

[5] http://resources.marshalladulteducation.org/

allowing for more than 200 total responses. 225 choices were selected altogether, 196 of them being the correct answer, resulting in 87.1% of the responses being correct.

Table 3-3. The proportion of correct responses selected by all respondents

Total Responses	Correct Responses	Correct Responses (%)
225	196	87.1%

In a perfect case, all 100 of the targets would be chosen, and all 300 of the distractors would not, resulting in exactly one fitting answer for each question. The distribution of correct and incorrect answers for each target and distractor is displayed in Table 3-4.

Table 3-4. The percentage of connectives questions deemed valid, invalid, and questionable

	Target valid	Distractor valid	Distractor invalid	Distractor questionable
All	100	300	300	300
Identified	96	273	4	23
% Identified	**96.0%**	**91.0%**	**1.3%**	**7.7%**
Total	**92.3%**		**6.8%**	

96% of all targets were identified by both respondents as being the correct answer, and 91% of distractors were identified as being incorrect. Of the 300 distractors, only 4 were identified as being fitting answers by both respondents, and 23 were identified as fitting by only one of the two respondents, making them questionable. Combined, this resulted in 92.3% valid choices, and 6.8% either invalid or questionably valid. These results suggest that the current method of selecting connectives from a passage and replacing them with random distractors of a different class is an effective method for

creating valid questions in most cases, but further curating is necessary to avoid introducing false-negative distractors.

3.3.3 Limitations and Future Work

The results in the previous section show that over 90% of the distractors generated were found to be valid. To examine the invalid and questionable distractors more closely, Table 3-5 breaks down the connective classes of each target-distractor pair to determine potential patterns of which connective classes were found to function interchangeably in the same context. Note that there is no distinction made between which choice was the target and which was the distractor, because both choices were deemed to be equally valid for the context in which they were presented.

Table 3-5. The distribution of connective classes that were deemed to be interchangeable when included as choices for the same question

	Temporal	**Contingency**	**Comparison**	**Expansion**
Temporal	12%	-	-	-
Contingency	24%	4%	-	-
Comparison	28%	12%	0%	-
Expansion	8%	8%	4%	0%
Total	**72%**	**24%**	**4%**	**0%**

The results show that the vast majority of questions with more than one fitting choice included a temporal connective: roughly half of all such questions contained a temporal connective alongside either a contingency or comparison connective, while in some cases, two temporal connectives of different subtypes (i.e. synchronous and asynchronous) were deemed to fit in the same context. In several cases, a comparison

connective and a contingency connective were found to be equally fitting. The only combinations that were not found to be interchangeable in any context encountered were two comparison connectives and two expansion connectives with different subtypes.

These results suggest that temporal connectives are likely to be the most problematic for being interchangeable with other classes, meaning that questions that test knowledge of temporal connectives need to take extra care when selecting distractors that are sufficiently different in meaning. Future work will further explore this phenomenon to examine when and how temporal connectives are likely to be interchangeable with other connectives, and whether certain connective classes and subtypes are more likely culprits. By determining this, future versions of the algorithms will be better equipped to generate distractors that do not fit in the desired context, improving the validity of the results.

Finally, as has been discussed in previous sections, future work should seek to prove the efficacy of these types of exercises for targeting inference-making skills in low-literate readers, and to determine whether the reading level or content of a given passage has an effect on the validity of the output.

3.4 Summary

This chapter described two unique reading comprehension question generation algorithms, each of which was designed to target a reader's ability to draw inferences and identify inconsistencies within a text. The first system, an algorithm for introducing locally-consistent inconsistencies, was found to be effective in many cases for replacing words in a text with words that both make sense in a narrow context and do not make sense in the full context of the passage. The results showed that the Google Books *n*-grams corpus can be successfully applied in new ways to assist in the creation of

comprehension monitoring questions. The second system, an algorithm for testing a reader's ability to apply discourse connectives, was found to be even more successful: human evaluations showed that more than 90% of the items generated were valid questions with only a single fitting answer. This chapter concludes with the assertion that it is possible for automatic generation systems to create useful and high-quality exercises for testing a reader's ability to monitor their own comprehension and draw inferences about a text by utilizing existing data sources in novel ways.

Chapter 4 - Application Design

This chapter discusses the design of a smartphone application for distributing the materials generated by the previously-described algorithms to adult learners. Section 4.1 begins with a brief discussion of the current state of the art in both language learning software and software interface design for illiterate users. Section 4.2 describes the backend structure of the system for distributing the materials to users, including the organizational structure of how materials are stored and the formats of all the practice assessments that the software supports. Section 4.3 describes the specific design choices applied to the application interface, including guidelines around user experience and the incorporation of the science of learning. Section 4.4 describes the design of an initial prototype application and the results of a think-aloud usability study with low-literate adult learners. Section 4.5 discusses the changes and additions made to the full application, and the final evaluation of this version of the app by both students and instructors. Finally, Section 4.6 concludes with a brief summary of the outcome of the software design phase.

4.1 Previous Work

This thesis is far from the first study to discuss the creation of a software application for language education. In fact, the field of Computer Assisted Language Learning, or CALL, has been prominently researched since the early 1980s, encompassing a wide array of different technologies, from word processors and web browsers to educational applications. Educational CALL systems have been shown to be effective methods for targeting elements of language acquisition that non-native speakers struggle with, such as

vocabulary knowledge [101-103] and grammar [104, 105]. However, the majority of CALL systems are focused on *language* learning (primarily second languages), assuming that the learner possesses a functional understanding of the basics of written and spoken language in their own native language, which cannot be assumed of low-literate learners.

Software for *literacy* learning is nearly always designed for children, and is primarily studied for its effectiveness compared to traditional intervention methods. Many studies have demonstrated the significant positive impact of computer-assisted intervention for improving literacy in children, including preschool-aged beginning readers [106], children with reading disabilities [107], and children with autism [108]. The vast majority of studies on the effectiveness of such software have examined the benefits for phonological and phonemic awareness, nearly all of which discovered significant performance gains for those children using the software [109-112]. Other studies have demonstrated similar performance gains from software interventions targeting specific skills such as letter identification [113], vocabulary [114], writing and meaning synthesis [115], and simple reading comprehension [116].

While there is much research to suggest that computer-assisted intervention can be an effective tool for bolstering literacy skills in younger learners, few software applications have been designed explicitly for functionally illiterate adults. In fact, [62] found that only half of the technologies currently available in adult education programs today are designed with adult learners in mind; the rest are either designed for K-12 learners, or are devoid of educational features altogether. Much of the adult literacy software that does exist is designed to provide text-scaffolding and other assistive tools for users with reading difficulties. In 1990, [117] discussed the design of a word processing program to

assist low-literate adults with grammar, spelling, and word meanings. It was designed with the specific needs and difficulties of the adult user in mind, with a simple interface, minimal navigation, and a dictionary with easy-to-understand definitions. More recently, [118] developed a robust dictionary-like application to provide assistive language support tools, allowing users to look up unfamiliar words and listen to their pronunciation and definition. [119] explored the development of a tool for enhancing and simplifying text found on the web, using Natural Language Processing tools to automatically simplify, shorten, and add elaboration to texts to adapt them for less skilled readers.

Other software, like this project, is meant to serve as a learning tool. [120] describe the creation of several tablet applications designed for low-literacy adults, with a set of exercises for punctuation placement, and commonly-confused homophones. The study explored the value of gamification in educational software for adults, finding that incorporating visual goals and rewards increased learner engagement. [121] explored the potential of SMS (i.e. text messaging) as a conduit for delivering micro-lessons and interactive reading quizzes to adult learners, finding substantial improvements in the reading level of participating users over a relatively short period of time.

Perhaps the most related tool to the proposed CAPITAL system is an Intelligent Tutoring System called ReadOn [122]. Designed to improve reading comprehension for those with intermediate literacy skills, the project combines a learner interface of passages and questions with an authoring tool for instructors to input and customize materials. The system keeps track of every student's individual progress, strengths, and weaknesses, and presents materials to them accordingly. However, while the system does have the capacity to automatically generate basic vocabulary questions and anaphora

resolution tasks from a passage, it largely relies on handcrafted and manually-categorized materials. Target users responded favorably to the software, although no conclusions could be drawn about its usability or the long-term benefits of its use.

Few of these studies outlined above describe in detail the design of the software and the ways in which the user interface was tailored to accommodate low-literate users. A number previous case studies have explored solutions to the challenge of designing software to assist illiterate populations in greater depth, but the vast majority of these differ from this project in two distinct ways. First, nearly all are designed to assist with concrete physical tasks such as navigation [123], banking [124], healthcare management [125], and even text messaging [126]. Second, most have explored how to design usable technology specifically for illiterate populations in developing countries where illiteracy rates are exponentially higher, being especially mindful of the unique cultural considerations for these specific groups [127-129].

CAPITAL differs from these in two notable ways. First, the target users for CAPITAL are native English speakers in the United States, which means that this software is not constrained by the need for culturally-agnostic iconography. Second, this software is educational in nature and designed to *improve* the user's literacy skills, not simply accommodate their absence. As such, this software faces a unique challenge in conveying abstract or higher-level learning concepts in a way that is still intuitive and easily-accessible for low-literate users.

4.2 Building the System

Items generated by the algorithms described in the previous chapters already have an implicit ordering which stems from their ascribed difficulty level (either from their

phonetic skills or their originating text difficulty). This is designed to ensure that students progress through the materials beginning with the easiest and working up to the more difficult items. However, because so many different items can be generated that target so many different skills, it is important to organize these materials so that students can access them quickly, easily, and intuitively.

The system organizes sets of questions into learning units called *exercises*, which are sequentially arranged within *courses*. Students are assigned materials at the Course level, where each course contains one or more exercises made up of items organized by increasing difficulty level. The items within each exercise are presented to the student sequentially in the order specified. Figure 4-1 shows a simple diagram of the hierarchical structure of these components.

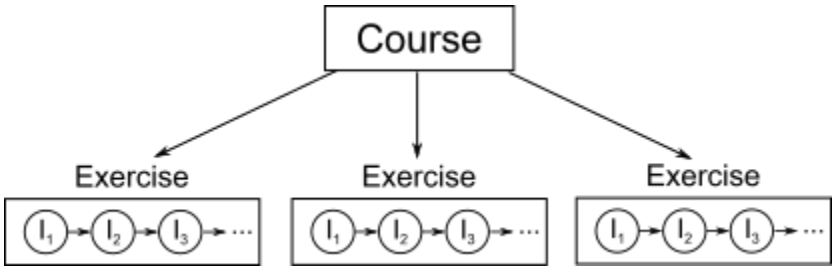


Figure 4-1. The hierarchy of the distribution system: Each Course holds a collection of one or more Exercises, each of which serves as a container for generated Items

To make them useful for self-guided learning without the supervision of an instructor, exercises must be supplemented with additional features. Exercises take the form of miniature assessments, through which the student can be presented with several different possible options for an answer to a given exercise and have the student select the correct choice to demonstrate their mastery of a concept.

Phonological awareness and word analysis exercises both target sounds within a word. As such, meaningful sounds must be provided to the learner. To supplement the transcribed phonemic breakdowns of the words in the system, the system includes an audio file for each word, obtained from the Google dictionary. For exercises which test basic word understanding, many words are also accompanied by a clear representative image which were hand-selected from OpenClipart.org [130].

The majority of the exercises take the form of multiple choice questions. For these exercises, the algorithms described in the previous chapters can be used to generate intelligent distractors for each target word, according to the specific skill being tested. The following is a brief description of each of the assessment types included in the application.

Find the Rhyme. These items target the user's ability to identify rhyming words by sound. Users are given the mp3 for a word and are asked to choose the mp3 of the word that rhymes with it. The basis is the mp3 that rhymes with the target, and each distractor is a word of the same difficulty level that does not rhyme with the target.

Sound It Out. These items target the user's ability to segment a word into phonemes. Users are given the mp3 for a word and are asked to choose the mp3 of a phoneme it contains. The basis is the mp3 of a phoneme found in the target and each distractor is a phoneme of the same manner of articulation as the basis that is not found in the target.

Pick the Word. These items target encoding skills. Users are given the mp3 for a word and are asked to select its written form. The basis is the word that matches the given mp3, and each distractor is a word that differs from the basis by a single phoneme.

Pick the Sound. These items target decoding skills and are the reverse of Pick the Word exercises. Users are given a written word and are asked to select its matching mp3. The basis is the mp3 that matches the given word, and each distractor is a word that differs from the basis by a single phoneme.

Spell the Word. These items target encoding skills. Users are given the mp3 of a word and a set of letter tiles which they must place in the proper order to spell the word they hear. Distractor letters are randomly chosen from letters that share visual and/or phonetic characteristics with the basis letters.

Pick the Wrong Spelling. These items target encoding and decoding skills, as well as contextual awareness. Users are shown the text of a sentence, with one of the words misspelled. Users must locate and physically tap on the misspelled word in the text.

What is This? These items target encoding skills as well as image identification. Users are shown an image and asked to select the written word that best represents it, and each distractor word differs from the correct word by a single phoneme.

4.3 Design Guidelines

One of the biggest challenges in designing an application for low-literate users lies in making the application easily usable without the need for reading. The software must be designed in a very deliberate and thoughtful way to ensure that it is simple, intuitive, and accessible to the target population. [131] found that the difficulties faced by functionally illiterate adults extend beyond strictly literacy-centric skills: adults with minimal reading abilities also tend to perform significantly worse in tasks requiring visual memory, spatial awareness, cognitive processing speed, and focus. With this in mind, the CAPITAL interface was designed around several key guidelines for creating usable educational

software for low-literate users. Each of these rules and the rationale behind them is described below.

1. Replace text with established iconography. An obvious choice when designing software for users with limited literacy is to minimize the amount of reading the user is required to do in order to navigate. Nearly every application that has been designed specifically for low-literate users has identified the need to minimize or eliminate text from the interface [123, 132, 133]. As with other studies such as [124, 127, 133, 134], the CAPITAL software instead relies on images and intuitive design to suggest to users what they can and cannot do within a given context.

Wherever possible, the design leverages common iconography that a user would likely already be familiar with: for example, left and right directional arrows represent moving forward and backward through the screens. However, low-literate adults have been shown to be more successful at identifying lifelike and colorized images as compared to black and white or cartoon-like iconography [131], requiring us to avoid overly abstract or minimalist picture representations of items.

This introduces a significant challenge when attempting to guide the user through educational activities, as many of the learning concepts are abstract in nature and cannot be easily represented through imagery. If a concept does not already have an intuitive pictorial representation that the user would recognize, the software attempts to build such an association organically through consistent repeat exposure. Unique icons are established to represent each item format, providing consistent visual clues to allow users to easily identify their upcoming tasks.

2. Avoid hierarchical menus. Navigation of a software interface requires basic spatial orientation and the ability to multitask and seamlessly shift focus and attention between different contexts [135], skills that functionally illiterate adults have been shown to struggle with [131]. This necessitates a simplified navigational structure, one which minimizes the user's need to rely on spatial awareness or visual memory to move between contexts. Following the recommendations of previous studies [127, 136], a linear structure is utilized rather than relying on hierarchical navigation. Wherever possible, user input is minimized to a semi-persistent "forward arrow" button which advances them to the next screen, allowing users to simply move forward screen by screen, with the last step forward returning them back to the dashboard as appropriate.

3. Incorporate whitespace. Because low-literate users have been found to have a narrower field of view when presented with information [137], screens are structured so that users do not become overwhelmed by too much information at one time, dividing content across sequential screens where appropriate to minimize content overload and spreading the remaining content out to fill the screen. Studies have also shown that illiterate adults tend to have weaker-than-average fine motor control [138], so to minimize a user's chances of unintentional mis-clicking, interactive components are designed to be large and prominent, with significant negative space between tappable elements.

4. Make related components easily distinguishable. Because CAPITAL is an educational application, the primary goal of its design is to minimize the cognitive load required for the user to actually reach their learning materials. This is accomplished through the use of consistent and distinct colors and effects to clearly imply element

functions and relationships, prioritizing a high-contrast and functional design over modern minimalist aesthetics.

Every interactive element is styled using the same consistent effects such as drop shadows and glowing borders to help distinguish them from their static surroundings. This allows users to easily recognize which components are of the highest priority and require their input, minimizing the chances of the user becoming confused about how to proceed. Colors are assigned with purpose and are only used for their assigned context: for example, each individual Exercise format is assigned its own unique color, which is carried through to every component related to said Exercise, including backgrounds, icons, and buttons. These consistent stylings are used to tie related components together and serve as visual cues for the functions of a given screen in the absence of written guidance.

5. Minimize barriers to entry. It is imperative that the software have as few barriers to entry as possible, to ensure that users do not become overwhelmed or confused before being able to access their materials. Perhaps the biggest barrier to entry for any system that requires a user account is the registration and login process. Typical applications require users to create a unique username or to register with an email address; however, many low-literate users do not use email services, and creating and inputting a unique, memorable alphanumeric username requires a base level of reading and writing ability that cannot be assumed for the users of this system.

To simplify this process, users are asked to register for an account using only their phone number. The system then assigns them a randomly-generated username, which is used only for public social features and for instructors who wish to monitor student

activity. Most importantly, the app stores the user's login information to their device so that they only ever need to log in once. Every subsequent time the user loads the app on this same device, the system will log them in automatically and forward them directly to their dashboard.

4.3.1 The Science of Learning

Recent studies have shown that there are three key features that enable effective learning: providing immediate feedback, exposing the learner to exercises slightly ahead of their current skill level, and distributing materials to them over time [143].

Providing immediate feedback is a largely trivial task for educational software, and is in fact one of the greatest strengths of dynamic software over static written materials as a learning tool. When a user submits an answer, the system is able to immediately tell them whether their answer was right or wrong, and to highlight what the expected answer was, ensuring that the proper knowledge is reinforced. Audio-visual cues are also incorporated, using a red and green color scheme to highlight wrong and right answers, and playing a cheerful-sounding jingle when correct responses are submitted.

A placement test system addresses the second component of effective learning: exposing the user to materials slightly more advanced than their current competency level. The placement test system ensures that students who are at a more advanced level are not required to work through materials that are too simple for them, which would otherwise feel like a chore and reduce engagement. Users begin each exercise with a placement test to determine their starting point. In a placement test, the user is shown a stream of items of gradually increasing difficulty, and the user's starting level is set to the level of the last item they answered correctly when the test ends.

To distribute materials to users over time, items are automatically selected for each user based on their performance, removing the need for them to decide their own path forward. Items are delivered to students in "rounds" of 10 at a time, and any item that is answered incorrectly in a round is reincorporated into the end of the round queue, requiring the user to answer every item correctly once before the round ends. If the user answers 80% or more of these items correctly the first time, they will unlock five new items, gradually introducing harder materials into the queue. Items that were answered incorrectly when the user last saw them are given highest priority in the queue, while the remaining items are chosen based on how long ago each was last seen and how many times the user has answered them wrong.

Additionally, the software incorporates two basic forms of gamification to entice learners to practice, as gamification has been proven to be an effective motivational tool for encouraging regular use [142]. The first is a simple point-accumulation system, where users are awarded points for completing exercises. More correct answers earn the user more points, and every user's cumulative point score allows them to be ranked in a public leaderboard. The second is a content-leveling system, where users are awarded badges to signify their increase in level when they advance to harder content. Both of these gamification practices are designed to give the user a sense of accomplishment and to clearly illustrate their forward progression. Finally, a 5-star meter at the end of each question round rewards users with a visual indicator of their short-term performance.

4.3.2 Cambourne's Conditions of Learning

The Conditions of Learning as theorized by Cambourne [144] describes eight individual conditions that are necessary to facilitate literacy development. A study of

digital textbook software explored how tablets and touch devices are an ideal platform for facilitating each of these learning conditions [145]. Inspired by this, the CAPITAL software has been designed to ensure that each of these learning conditions is met. The following is a breakdown of each of Cambourne's eight conditions and how they are realized in CAPITAL.

Immersion. Learners need to be immersed in their learning with intellectual and sensory stimuli. The exercises strongly emphasize the relationship between writing and sound, and all of them require active engagement through touch interaction.

Demonstration. Learners must be given practical demonstrations of what is being learned. Each exercise begins with an animated tutorial for how to answer each item, which users then mirror through their own interactions.

Expectations. Learners must be given clear indicators of what is expected of them. Progress bars are displayed inside exercises to show how far the user is to completion, as well as level meters that fill up over time, and blinking effects for exercises that users should most focus on.

Engagement. Learners must feel actively engaged with the materials they are learning. The system inherently accomplishes this with a learn-by-doing approach, since users are active participants rather than passive observers.

Responsibility. Learners must feel in control and personally accountable for their learning. Users are given full control over their learning path, but also encouraged to do the work that the system determines is most valuable for them with Daily Exercises and login streaks.

Approximation. Learners must be allowed to make mistakes without fear of punishment. The reward system has thus been designed to give points for correct answers while not detracting for wrong ones. The user is also able to redo their missed items at the end of a round.

Employment. Learners must be given time to interact with the content in realistic ways. Items are delivered in small rounds at a time to avoid overwhelming the user with lengthy expectations, and they are encouraged to prioritize their assigned Daily Exercises if they are short on time.

Response. Learners must be given timely feedback. The immediate feedback system ensures that learners are always immediately informed of their performance on an item.

4.4 Initial Prototype Design

A prototype version of the CAPITAL application was created to evaluate the efficacy of the general design as a usable tool for low-literate users. The prototype version was intentionally designed to be as simple as possible, lacking many of the "flashier" features that the full version would include such as the leveling system and performance-based item distribution. The focus instead was placed upon the effectiveness of the iconography, the navigational structure, and the overall accessibility of the software.

The prototype version contained only one type of exercise, where the user was given an mp3 and asked to select the written word that matched it from a list. A card-based interface displayed the interactable content, presenting course cards in a scrolling vertical list, each represented by a unique image. Inside each course, exercises were displayed in a horizontal carousel of cards, where one exercise had to be completed before the next

was unlocked. The most recently unlocked exercise took focus by default, allowing the user to simply press “Go” to move forward.

Users were shown clear and obvious cues that immediately informed them of whether they selected the right or wrong answer, both visually and audibly. The correct answer was also prominently emphasized at the top of the screen so that the user could easily compare against their own answer and learn what mistakes, if any, were made. Finally, a simple 5-star meter was presented at the end of every exercise to show the user’s overall performance.

4.4.1 Student Think-Aloud Evaluation

To test the prototype app’s usability, 11 adult learners from the Washington Literacy Center (WLC) in Washington, D.C. were invited to test the software in a guided study. IRB exemption was obtained for this study, but all subjects were willing volunteers and maintained anonymity throughout. 7 participants were men and 4 were women, and all were at the lowest level in the WLC curriculum. 7 students owned smartphones or were familiar with their use, and 4 had never used a smartphone before.

Prior to beginning, each student was shown a 5-minute narrated video outlining the purpose of the app, the concepts of courses and exercises, and how to think aloud for the study. The participants were intentionally not told how to use the app or complete any of the tasks.

Each student was given a unique user account enrolled in the same four courses: the target course and three “decoys.” The target course contained three exercises with five items in each. Students were asked to enter the target course, complete the first two exercises, and then retake the first exercise again.

During this process, each user's successes and failures were recorded over the course of completing 10 different tasks. Four tasks were navigation-based: scrolling horizontally, scrolling vertically, entering a course, and selecting a radio button. The remaining six tasks were identification-based: recognizing the meaning and/or function of the target course image, speaker icon, arrow icon, "Go" button, check/x-marks, and star meter). Successes and failures were identified by observing each user's deliberate actions; unintended actions, such as accidentally tapping a button when scrolling, were not recorded. Verbalized thoughts were also included in the failure counts, such as a student saying, "I think this is it" while gesturing to a decoy course.

4.4.2 Results

Errors were recorded in the context of two different metrics: how *intuitive* the software was, and how *learnable* it was. An intuitive feature is one that users can quickly discern without the need for extensive trial and error, while a learnable feature is one that can be easily remembered on subsequent interactions. The intuitiveness and learnability of each task was evaluated by examining the number of errors that users committed before their first successful attempt and after their first successful attempt, respectively.

Figure 4-2 and Figure 4-3 show the distribution of failed attempts made before and after the first successful attempt, for navigation tasks and identification tasks, respectively. The results show that users with no prior smartphone experience struggled to complete more tasks than users who owned smartphones, particularly with navigational tasks such as scrolling and tapping. Horizontal and vertical scrolling proved to be the least intuitive navigation-based tasks for both groups, though they were both largely learnable for all users. Most users immediately understood how to interact with

radio buttons, though some inexperienced users struggled with the actual act of tapping. Several users struggled with figuring out how to tap a course card to enter, and this also proved to be the least learnable task for inexperienced users, though experienced users learned this easily.

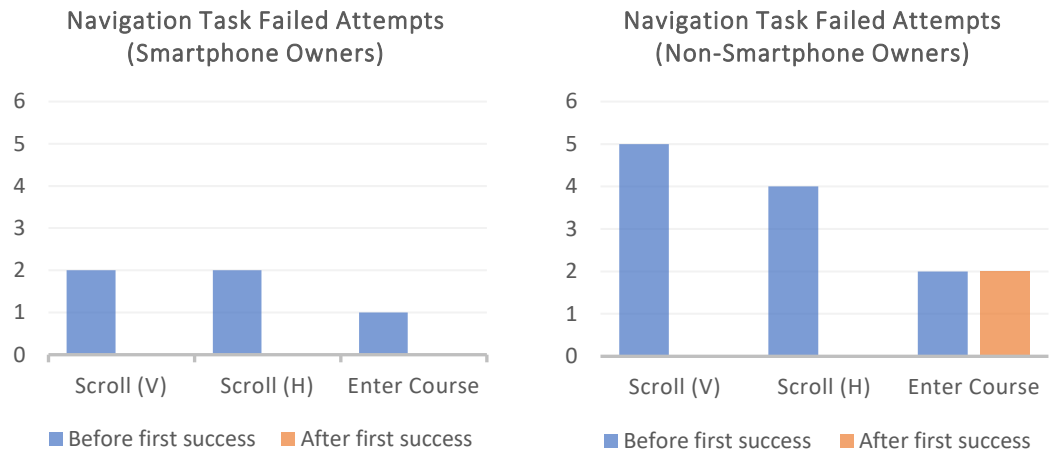


Figure 4-2. Failed attempts occurring before and after the first successful attempt for navigation tasks (smartphone owners vs. non-smartphone owners)

The only component immediately identifiable to all users was the "Go" button, although experienced users easily recognized most components. Locating the target Course proved to be the least intuitive task for all users; despite being shown the target Course's image prior to beginning the study, more than half of the users failed to locate it on their first try. However, after successfully locating the Course once, nearly all users were able to do so again with no errors, suggesting that the association is learnable through repeated exposure.

Following the test, each student was also given an anonymous survey about their overall feelings using the app. Instructors administered the survey to each student in a private setting to encourage honest responses. Each prompt was read aloud to the student,

and responses were given verbally following a 5-point scale from “strongly disagree” to “strongly agree”.

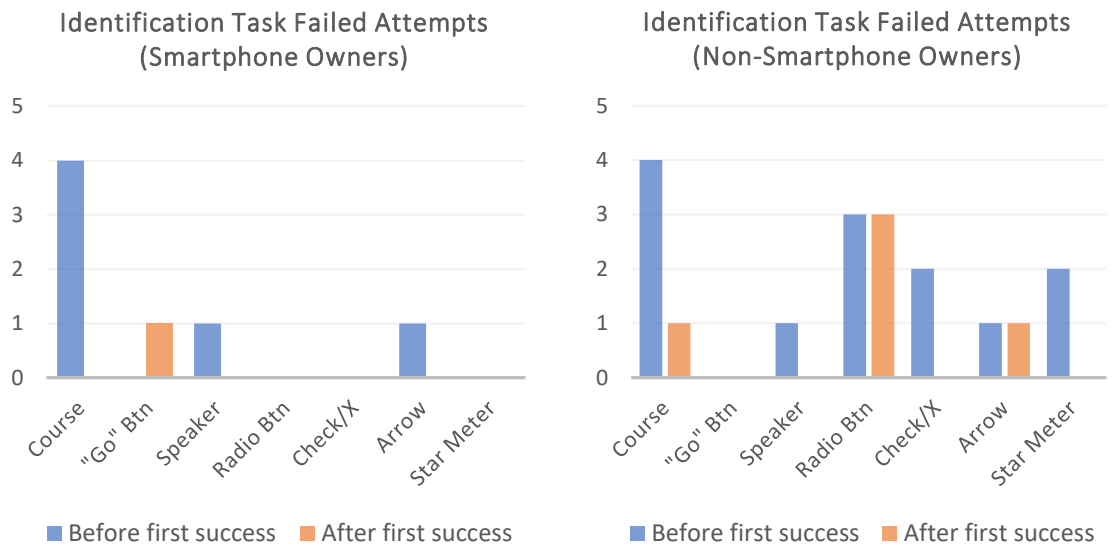


Figure 4-3. Failed attempts occurring before and after the first successful attempt for identification tasks (smartphone owners vs. non-smartphone owners)

The mean results of the opinion survey can be seen in Table 4-1. Several questions were repeated with opposite wording to ensure that participants understood the meaning behind their rating: for example, respondents were first asked if they found the app “easy to use” and later asked if it was “hard to use”. Three students’ responses were omitted because they answered such questions in a contradictory manner (i.e. responding “Strongly Agree” for both).

Despite the difficulties faced during the usability test, students responded very positively to the app. Students overwhelmingly agreed that it was enjoyable and easy to use, and most agreed that they would be able to keep using it independently.

Table 4-1. Average student responses to the opinion survey on a 5-point scale

	Question	Avg response
+	I would like to use it often	4.8
+	It was easy to use	4.1
+	I can use it without help	3.8
+	It is easy to find where to go	4.6
+	It will be easy for most people to use it	4.8
-	I needed to learn a lot before I could use it	2.8
-	It was hard to use	1.6
-	I can't use it without help	2.1
-	I need more help before I can use it	2.1

4.5 Final Design

Although the majority of the tasks evaluated in the prototype application were shown to be learnable over time, several tasks were not very intuitive for first-time users, particularly horizontal scrolling and identifying and entering a target course. Though respondents largely agreed that the software was easy and enjoyable to use, many were less confident that they would be able to use it without help.

The goal was then to address these issues in the final design to make the interface more intuitive and to increase students' feelings of self-sufficiency when using the app independently. In addition to the full set of features described in the previous sections, the following changes were also made to address the difficulties of the prototype:

Addition of audio and visual help. Optional audio assistance was added on every screen to help guide users who are having difficulty. Studies have shown that audio feedback significantly improves understanding in low-literate users [123], especially when coupled with visual animations [141]. Because low-literate adults face difficulties in processing language, verbal instructions are formulated to be as simple and succinct as

possible, and to provide the user with the ability to play audio back as many times as needed. The help button is persistent in the top right corner, and clicking on this button will provide audio descriptions of what can be done on the current screen. For the different exercise assessment types, the user is also presented with visual illustrations to accompany the audio instructions, showing them how to interact with every exercise and how to submit their answers.

Minimize or eliminate scrolling. The think-aloud study clearly showed that both competent smartphone users and those who were unfamiliar with mobile devices struggled to understand how to interact with the horizontally-scrolling Exercise list, and unfamiliar users also had difficulty with the vertical scrolling. To simplify the design, the Exercise carousel was removed altogether; the system was modified to instead dynamically choose which group of questions to present to the user within the given Exercise and present those questions automatically, with no need for additional navigation or user input. Additionally, the vertical list of Courses was replaced with a grid, maximizing the user's opportunity to see all the available material with minimal scrolling required (see Figure 4-4).

Evaluation of the final version of the CAPITAL app was done in two stages. First, a heuristic evaluation was conducted, where knowledgeable experts in adult literacy education were asked to give their feedback on the app's design and functionality and its potential effectiveness as a learning tool for low-literate adult students. Second, a small group of students was asked to use the final version of the app and participate in a second think-aloud study, wherein each student was asked to perform a series of tasks and their

successes and failures were recorded. The following sections describe these evaluation methods and their results.

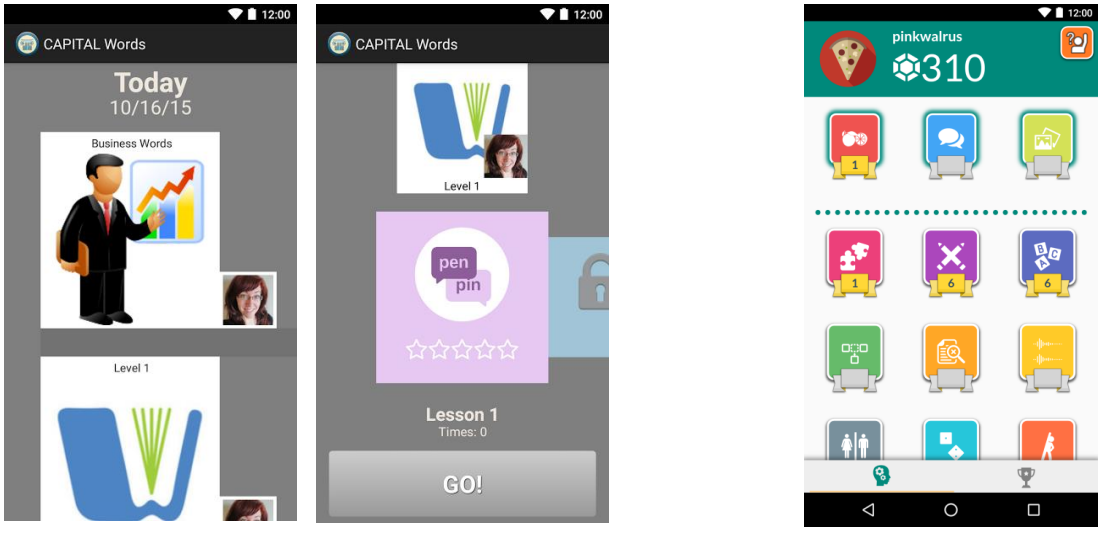


Figure 4-4. The vertically-scrolling list of course cards and horizontal exercise carousel (left) were replaced with a grid of exercises, within which questions would be dynamically selected for the user (right)

4.5.1 Instructor Heuristic Evaluation

Five adult literacy instructors were employed for a heuristic evaluation, with teaching experience ranging from 3 to 35 years. All five instructors owned and were familiar with smartphones, with all but one describing themselves as “comfortable” or “very comfortable” using mobile devices. To conduct the survey, instructors were given access to the app and asked to familiarize themselves with all of its features before providing their feedback on 62 statements on a 5-point Likert scale (from “Strongly Disagree” to “Strongly Agree”, respectively). Each statement described a feature that would be present within a well-designed application, making 5 the ideal score for every statement.

Instructors were also given the opportunity to explain their choices in each section and to provide any additional feedback not addressed by the survey itself.

The questions in this survey were adapted from a published heuristic framework known as MUUX-E, a comprehensive literature-based instrument specifically designed for evaluating m-learning software [3]. The previously-published evaluation of this framework supports the validity of the survey questions employed in our own study; however, due to the small size of the target population and the difficulty in obtaining participants, no attempts were made to ensure the reliability of these same questions, a shortcoming which should be addressed in future work.

The MUUX-E framework encompasses five evaluation categories: user experience, general interface usability, web-based learning, educational usability, and mobile learning. Each category is divided into several sub-criteria, each of which targets a specific component of successful software design.

Table 4-2 summarizes the average responses of all instructors within each category and subcategory. The average scores for all 5 categories ranged from 4.3 to 4.7, with no subcategories receiving an average score of less than 4, indicating significantly positive reception to all aspects of the software. The subcategories that received the lowest average scores related to error prevention and recovery, feedback, and ease of use as a system, while the highest-scoring features concerned the software's match to the real world, consistency, flexibility as a mobile tool, and minimalist aesthetic.

One instructor described their dissatisfaction with the depth of the feedback provided: "When a wrong answer is recorded, it's possible to find the right answer, but the app does not explain why the answer was wrong. The students will still have to find another way to

really understand the rules or principles that are followed." A future version of the app might find a way to incorporate more detailed feedback about why an answer is right or wrong based on the characteristics of the target and its distractors. Another instructor found the answer submission process somewhat unintuitive, stating that she "was a little confused when [she] didn't get feedback about the answer until [she] hit the arrow." She said that she found it "a little odd" that the same arrow button was used to both submit an answer and to continue to the next question. This is unfortunately a difficult design element to replace, as some choices include audio that require the user to tap on the choice itself to hear it, meaning that the choice selection must be submitted using a different form of input. When asked for suggestions on how to make the process more intuitive, the instructor in question was unable to come up with a better alternative.

Table 4-2. Average responses in each category and subcategory of the instructor survey

Category	Score
User Experience	**4.6**
Emotional issues	4.7
Contextual factors	4.7
User-centricity	4.4
Appeal	4.8
Satisfaction	4.7
General Interface Usability	**4.4**
Visibility of system status	4.0
Match to the real world	4.8
Learner control and freedom	4.4
Consistency	5.0
Prevention of usability errors	4.1
Recognition (vs. recall)	5.0
Aesthetics and minimalism	4.9
Recovery from errors	4.1
Help and documentation	4.2

Category	Score
Web-based Learning	**4.4**
Simple, well-organized navigation	4.4
Relevant pedagogical content	4.7
Suitable content of high quality	4.6
Easy to use system	4.0
Educational Usability	**4.3**
Clarity of goals/objectives/outcomes	4.3
Error recognition/diagnosis/recovery	4.2
Feedback, guidance, and assessment	4.1
Mobile Learning	**4.7**
Handheld devices and technology	4.6
Flexibility	4.9
Interactivity	4.5

On the other hand, one instructor was pleased with the way that the app dynamically selects material for the student: “I thought it was great that if you re-opened an app[sic] to do it again, it would start with new words.” Another instructor said that they “loved the fact” that audio prompts could be repeated as many times as needed by the user. One instructor who completed the survey provided only a single additional comment along with her answers: “This app is amazing!”

4.5.2 Student Think-Aloud Evaluation

A second think-aloud evaluation was conducted for the final version of the app, following the same protocol as the prototype evaluation. The goal for this evaluation was to ensure that the full-featured version of the app was still usable and appealing to the target users after the extensive redesign, and that the new features did not detract from the usability of the prototype version.

For this second round of evaluations, six new students were employed from the beginning reading level classes at Literacy Volunteers and Advocates (LVA) in Washington, D.C., ranging in age from 33 to 72 years old. As before, all subjects were willing volunteers and their anonymity was maintained throughout the study. 5 of the 6 participants were women, and all but one subject spoke English as their first language. Each subject was asked to rate their comfort with using smartphones and computers on a scale from 1 (very unfamiliar) to 5 (very familiar); although all six subjects owned a smartphone, only three subjects rated their comfort as a 4 or 5, and the other three rated themselves a 1 or 2, giving us two evenly-sized groups of “smartphone users” and “non-smartphone users” to compare.

Unlike in the prototype evaluation, subjects were not shown a video before beginning this evaluation, because the app already has a built-in tutorial video for new users upon registration. The video walks users step-by-step through every feature of the app and how to access them, using voice-over dialogue and illustrative images annotated with arrows and highlighted areas to point out buttons and their functions. As before, the test proctors did not explain to subjects how to use the app or complete any of the tasks, and each subject was regularly prompted to think aloud during the course of the study.

Each participant was asked to perform the same six tasks: navigation-based tasks asked participants to scroll to the last exercise in the grid, locate the leaderboard, and navigate back to the exercise dashboard, while identification-based tasks required participants to identify their curated daily exercises, the help button, and where they ranked in the leaderboard.

4.5.3 Results

Although some tasks were difficult for subjects at first and required significant trial and error, neither users nor non-users failed to complete any of the tasks after their first successful attempt. Thus, the graphs in Figure 4-5 compare the total number of failed attempts by both user groups before the correct solution was discovered, noting that no failures occurred after this time.

As was seen in the prototype evaluation, non-smartphone users tended to struggle more to complete most tasks, particularly those requiring navigation. Locating the bottom tab bar and discovering how to toggle between the Leaderboard and Exercise tabs proved troublesome for both users and non-users, possibly due to the muted colors and

inconspicuous placement at the bottom of the screen below the larger, brightly-colored components (see Figure 4-6).

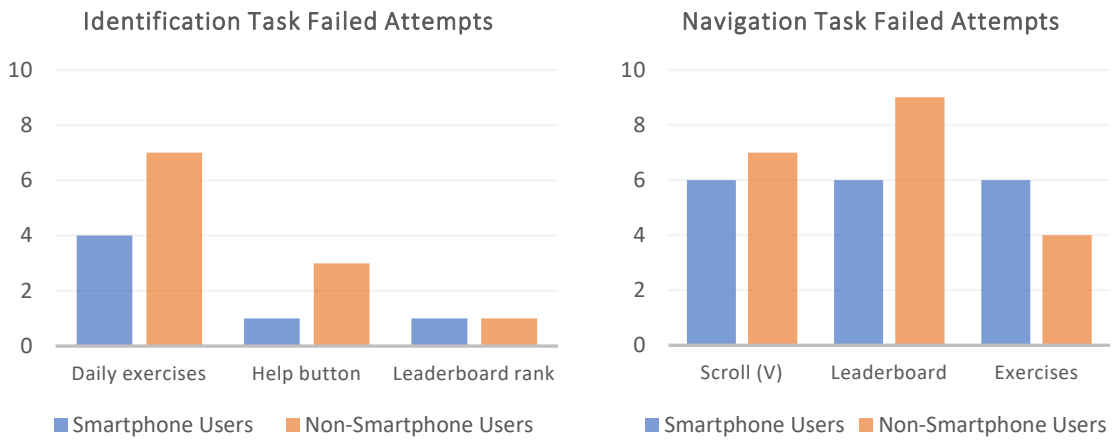


Figure 4-5. Failed attempts by smartphone owners and non-owners before the first successful attempt for identification tasks (top) and navigation tasks (bottom)

Beyond the tasks described in the charts above, there were several unexpected difficulties that subjects encountered during the test. All non-users experienced some difficulty in getting their button taps to register due to lack of familiarity with the necessary motion; by either pressing on the button for too long or accidentally dragging on release, their motions would not trigger the tap function, resulting in significant frustration. “This always happens,” one subject said after failing to activate a button several times in a row. “Phones don’t like me.” Another subject tried positioning the phone different ways, first holding it in their hand and then putting it flat on the table, but continued to struggle regardless of the screen’s orientation. This appears to be a problem specifically for users unfamiliar with smartphones, as no smartphone users encountered these difficulties.

It is interesting to note that several of the most failed tasks (identifying the daily exercises and toggling between the leaderboard and exercise views) were explicitly described in the tutorial video shown to users at the beginning of the evaluation. For example, the video shows an illustration of the exercise dashboard and highlights the blinking bordered exercises at the top while audio explains that "these exercises give you double points" (see Figure 4-7); however, when asked to identify which exercises gave double points immediately after the video ended, the majority of users could not identify them. One subject remembered that it had been explained in the video but immediately admitted, "I didn't even pay attention. Now I've got to go back and start all over."

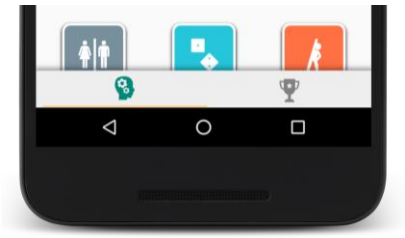
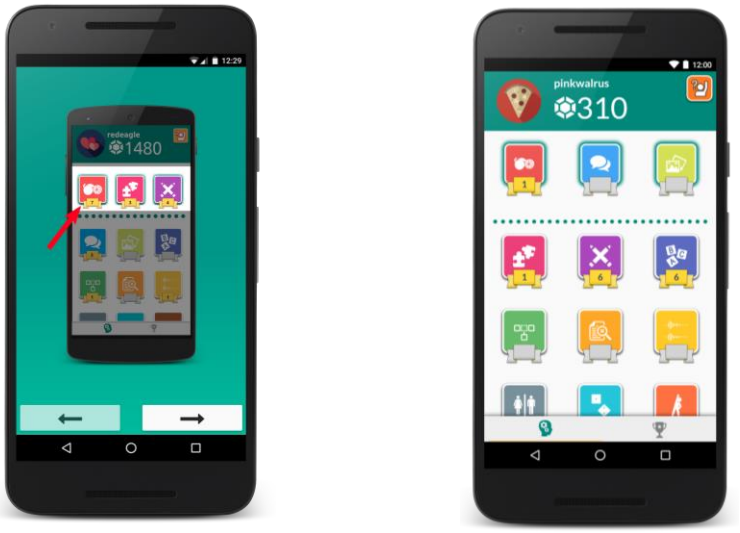


Figure 4-6. The tabs in the bottom bar proved difficult for users to locate and identify, despite being explicitly described in the tutorial video

Apart from forgetting the information presented in the video, subjects also faced confusion about the function of the narrated tutorial videos that appear before every exercise. These tutorials are designed to mimic what the actual exercise will look like and guide users on how to answer the questions using step-by-step audio and illustrations. Many of the tutorials begin with the instruction "Tap the speaker icon to hear a word" while showing an illustration of a finger hovering over a speaker button (see Figure 4-8); when presented with this combination of image and audio prompt, every subject attempted to tap the illustrated image of the button as if it were real. While the

smartphone users made this mistake only one or two times and quickly realized their error, non-users had significantly more trouble understanding: all of them tried tapping the image multiple times, and became confused about how to proceed when it did not produce a result. Additionally, two of the three non-users made the same mistake in at least one subsequent exercise tutorial, even after recognizing their error from the previous encounter.



(a) (b)

Figure 4-7. (a) The tutorial screen which describes the blinking exercises at the top that “give double points”; (b) The dashboard where subjects were asked to identify the exercises that would give them double points

This raises a very important question about the best way for software to convey instructions to low-literate users in the absence of written descriptions. The tutorials were included to help the users understand exactly what was expected of them by providing clear, visual, step-by-step instructions; however, these results suggest that static illustrations may be confusing or otherwise ineffective at clearly conveying concepts to

these users, and that consecutive instructions shown to users before they are needed can easily be forgotten by the time the user is expected to refer back to them.

Figure 4-8. When presented with this illustration and the instruction “tap the speaker icon to hear a word,” every subject tried to tap the illustrated button.

Despite these difficulties, all six users said that they enjoyed the app. Several students were extremely enthusiastic, expressing that it was “fun” and that they “loved it,” and all but one said that they thought it was going to help them learn, with the remaining student insisting that he needed to use it more first before coming to a conclusion. Regardless, every student requested to have it installed on their personal phones so that they could continue to use it once the test concluded.

4.6 Limitations and Future Work

The most obvious limitation when testing the usability of the application with real-world users comes from the relatively small sample sizes for the evaluations conducted. Ideally, evaluations would have been conducted with larger groups of students controlled

for a variety of variables, such as age, native language, smartphone experience, and reading proficiency level. Although numerous efforts were made to enlist students in multiple different ABE programs, success was severely limited by the number of students physically willing and able to attend sessions. The average class typically ranged in size from 3-5 students, and attendance was never guaranteed: for any attempted evaluation session, if 20 students were recruited, only 5-8 on average would actually attend. This made it extremely difficult to conduct sufficiently-sized evaluations, much less control for different variables in the subjects themselves.

Several follow-up evaluations were intended to be conducted. One evaluation would have asked the same students from the think-aloud sessions to return after one week to participate in a second evaluation with an identical task set; the results of this study would have given greater insight into the long-term learnability of the software rather than simply the short-term results of the first session. A second evaluation would have recruited a group of students to use the app independently over a period of several weeks, with regular follow-up meetings to learn their specific needs, desires, and difficulties. However, attempts to follow up with the same group of students over even two sessions were rarely successful. Because the subjects were evaluated anonymously, only the instructors were aware of the names and contact information of the participants, and were therefore responsible for attempting to regroup the same students for follow-up sessions; however, many students were physically unable to attend sessions with any consistency, and many others were difficult for instructors to reach by phone, making it extremely difficult to follow up with the same students for multiple sessions. These difficulties are

consistent with what is known about student attrition rates in ABE programs, and are extremely difficult to account for in studies of this nature.

Future work would greatly benefit from more extensive, long-term evaluations with a steady group of users. Although this thesis was constrained by the limited number of students we were able to recruit, we will continue to attempt to find a more consistent and sizeable group for future evaluations.

Additionally, more exploration should be done into how best to present instructions to low-literate users. As described in the previous section, users did not respond well to being instructed through voice-over tutorials, either ignoring or simply forgetting the information provided by the audio dialog. Additionally, still-frame narrated video guides were found to be confusing and easily misinterpreted, with users mistaking the tutorial images as components to interact with. Future work should seek to find a better method of illustrating and conveying instructions to users in a way that clearly conveys its function as a tutorial, and does not tax the user's short-term memory, challenge their recall, or otherwise disrupt their workflow.

4.7 Summary

This chapter discussed the design stages of the CAPITAL smartphone application for delivering exercises to students. The software was carefully designed to be as usable as possible by users with below-average literacy, following a series of specific design guidelines supported by the existing literature to ensure consistency, ease of use, and a sufficient amount of guidance provided without the need for reading. The application also incorporated the three key features identified to enable effective learning: providing immediate feedback, providing learning materials that are slightly ahead of the user's

current level, and delivering these materials over time based on past performance. Think-aloud evaluations with real adult students in literacy programs showed that the design of the application is relatively intuitive for users who are regular smartphone users, and highly learnable for all users, even those with minimal digital literacy. Expert instructors confirmed that the application is well-designed as a learning tool for their students, and the majority of student users expressed enthusiasm for wanting to continue to use the application after the evaluations concluded, feeling largely confident that it would help them learn and that they could use it without assistance. However, the results of user testing indicate that the average low-literate user does not respond well to illustrated tutorial videos, either forgetting the information presented in the videos or misinterpreting the video components as elements of the actual software. Future work will seek to determine how best to provide instructions to students in a way that minimizes cognitive load.

Chapter 5 - User Engagement

Following the usability studies, two cohorts of smartphone-owning student volunteers were given the app on their personal devices to use at their own discretion. 18 total students were given access to the app: 11 students from beginner and intermediate level classes at the Washington Literacy Center (WLC) and 7 beginner-level students from Literacy Volunteers and Advocates (LVA). The following sections report the usage data for each of these cohorts during their first 30 days of use.

Any user who used the app for at least 5 days was considered to be an "engaged" user. The majority of users with fewer than 5 days of activity used the app only on the day it was first installed, with one user logging in a second day but not answering any questions during that time. Of the 18 students monitored, only 4 used the app for less than 5 days, with 3 of the 4 not logging in again after the first installation. 9 of the 11 users from WLC and 5 of the 7 users from LVA were engaged.

Table 5-1 shows the breakdown of how often engaged users logged in and answered questions within their first 30 days of use. On average, engaged users used the app approximately 14 of the 30 days, and logged in more than twice a day, equating to multiple usages roughly every other day.

Table 5-1. The number of usage days and individual logins for all engaged app users in the first 30 days

	Usage days	**Total logins**	**Logins/day**
LVA	13.6	40.6	2.8
WLC	14.6	30.1	1.9
Total	**14.2**	**34.2**	**2.3**

Table 5-2 breaks down the number of questions answered by all engaged users in the first 30 days. These statistics include the average number of questions answered by each user, irrespective of whether the question had been answered before, compared against the number of *unique* questions answered per user. The results show that the average user submitted over 2,000 question responses in their 30 days of use, being exposed to more than 300 unique questions during that time. This means that, on average, each question was answered more than 5 times per user; because the system is designed to repeat questions for the user when they were not answered correctly, this suggests that it took users many attempts to learn the correct answer for a given question. This also shows that users were not put off by the fact that they were seeing the same questions multiple times, as they continued to use the app without being discouraged to stop logging in.

Table 5-2. The average number of unique questions answered, and the total number of question responses received, for all engaged users in the first 30 days

	Average
Questions answered	2077.8
Unique questions answered	333.2
Repeats per question	5.5

Finally, Table 5-3 shows the popularity of each of the assessment types, according to how many users chose to engage with them, and the average number of questions that each engaged user answered within. The average level gain for each user is also included, calculated by comparing the level that each user reached in the placement test to the level of the highest question he or she unlocked in the 30 day span.

The results show that the most popular exercise was the Spell It format, which asks users to listen to a word and arrange letter tiles to spell the word they hear. All but two

users chose to answer questions in the exercise, and the engaged users each answered approximately 800 questions in this exercise alone in just 30 days. The Pick the Word exercise engaged fewer users, but the users who were engaged answered close to 400 questions each. Other popular exercises were the Sound It Out, What Is This?, and Find the Rhyme exercises, each of which engaged at least 10 of the 14 users, who answered between 150-300 questions in each category on average. The least engaging exercises were Pick the Sound and Pick the Misspelling, which engaged only 4-5 users each; however, the users who were engaged with these exercises still answered over 130 questions each.

Table 5-3. The average level gain in each exercise for all engaged users in the first 30 days

	Sound It Out	**What Is This?**	**Find the Rhyme**	**Pick the Word**	**Pick the Sound**	**Spell It**	**Pick the Misspelling**
Users	11	10	10	8	5	12	4
Responses	276.3	162.9	207.9	398.5	138.8	792.7	139.0
Level Gain	1.1	1.3	1.2	2.0	0.8	0.5	0.6

Users experienced the greatest level gain in the Pick the Word exercise, increasing by two levels on average in the 30 day span. The Sound It Out, What Is This?, and Find the Rhyme exercises also saw an average gain of over 1 level per engaged user. However, users experienced at least a minor increase in level across all exercises in the 30 day span, unlocking harder questions than the ones they began with.

Although anecdotal, these results suggest that the app is engaging enough to grab and maintain students' interest over a sustained period of time, and that students feel motivated to use it independently as a learning tool. Students who enjoy the app will typically use it multiple times every couple days, and they are willing to explore different

exercises. Students also do not seem to be discouraged by the fact that they often have to repeat the same questions many times before unlocking more, likely due to the way the system presents questions in dynamic rounds each time. Students were also found to be able to make minor learning gains through regular use, unlocking harder materials after enough practice. Future studies will explore students' enthusiasm for the software over a longer period of time, and will seek to explore the app's potential as a learning tool more definitively compared to more traditional methods.

Chapter 6 - Conclusion

According to current population numbers, a remarkable 13 million adults in the United States alone are functionally illiterate, with an additional 150 million burdened by sub-par literacy. These adults face varying degrees of difficulty in their day-to-day life, lacking the reading skills necessary to get by without struggle in modern society.

This thesis sought to contribute to the efforts to address these staggering illiteracy rates in three ways, utilizing novel algorithms for automatic question generation in both alphabetics and reading comprehension, and discussing the development of a smartphone application for delivering these generated practice materials to adult learners.

The first contribution of this thesis was a customizable system for generating all permutations of alphabetics learning for any given curriculum. This is first work of its kind to address the challenge of automatic generation of learning materials at the level of phonological and alphabetic literacy. Generation algorithms were developed for finding words with shared features, such as rhymes, phonemes, and letter combinations. Algorithms were also developed to create misspellings for words to test different types of decoding failure. This thesis also described a novel method of letter-phoneme alignment for words with known phonetic pronunciations, using knowledge of orthographic and phonetic syllable boundaries to inform the alignment of letter clusters and their sounds. Finally, a method was discussed for constraining the generated output of the system to preserve the intentions of the curriculum from which the items were created, never introducing word pairings beyond a student's current proficiency level. The generators described in this paper were shown to produce learning materials that are valid, correct,

robust, and thorough, proving that such a system can achieve sufficient coverage of all alphabetics skills while maintaining a high quality of output.

The second contribution of this thesis explored the automatic generation of several forms of comprehension monitoring exercises, specifically designed to target a reader's ability to draw inferences. Each generation algorithm explored a new application of an existing data source. The first was a novel method for finding the most contextually relevant words in a text, and introducing deliberate inconsistencies in their place that require inferential skills to identify. This system explored a novel application of the Google Books *n*-grams corpus for choosing words to make sense within a narrow context when substituted for the original word, and human evaluations proved that it was relatively successful at producing valid inconsistencies. Also described was a novel application of a discourse parser for creating questions to challenge a reader's understanding of connectives, and more than 90% of the items generated by this algorithm were found to be valid. These are the first question generation systems of their kind to attempt to target these specific reading comprehension challenges and utilize these data sources in the ways described. The results of both algorithms strongly suggest that it is possible for such systems to create high-quality exercises for comprehension monitoring and inference making.

The final contribution of this thesis was a thorough exploration of the design of a smartphone application for adults with below average literacy. This is the first study of its kind to address the specific challenges of balancing educational mobile software design best practices with accessibility design for low-literate users. The paper describes a series of specific design guidelines supported by the existing literature that were employed in

the design of the CAPITAL system to ensure consistency and ease of use while also prioritizing learning as the primary goal. The application also incorporated the key features of effective learning, providing immediate feedback, choosing learning materials based on the user's current proficiency, and customizing the rate of distribution. Evaluations with students in literacy programs showed that the final design was relatively intuitive for users familiar with smartphones, and highly learnable for all users, even those with minimal digital literacy. Expert instructors confirmed that the application was well-designed as a learning tool, keeping in mind the unique needs of the target user base. This paper asserts that the design choices made for the CAPITAL smartphone application resulted in a usable and effective learning tool for adults with minimal literacy. Additionally, anecdotal observations of student usage over a period of 30 days showed that the majority of users were engaged with the software, using the app periodically and answering a variety of questions, and that many users were capable of unlocking harder materials through regular practice.

CAPITAL is the first software of its kind to address the adult literacy crisis by both allowing for easier creation of learning materials and by providing a more effective method of delivering them to students in need. The CAPITAL system allows students to continue following the same curricular structure as what they are learning in class, circumventing many of the inherent difficulties that students often face in attending physical classes. "You come here for a couple of hours," said one of the app's most engaged users, "but it's so hard. You forget." When asked why they were so eager to use the app, the student explained: "We don't have school for the next 30 days, and we really need to get this, so we'll be practicing [the exercises in the app] instead." By providing

students with the ability to practice anywhere and at any time with a learning experience customized to their individual needs, CAPITAL addresses a glaring need in adult basic education programs today, giving students the autonomy and flexibility they ultimately need to be successful.

References

[1] I. S. Kirsch, "The International Adult Literacy Survey (IALS): Understanding What Was Measured," *ETS Research Report Series,* vol. 2001, *(2),* pp. 61, 2001.

[2] B. D. Rampey *et al*, "Skills of U.S. unemployed, young, and older adults in sharper focus: Results from the program for the international assessment of adult competencies (PIAAC)," 2016.

[3] UNESCO, "Records of the General Conference," *Records of the General Conference,* vol. 20, *(October),* pp. 18-22, 1978.

[4] I. S. Kirsch and A. Jungeblut, *Literacy: Profiles of America's Young Adults.* Princeton, NJ: National Assessment of Educational Progress, 1986.

[5] J. Baer *et al*, "Basic Reading Skills and the Literacy of America's Least Literate Adults: Results from the 2003 National Assessment of Adult Literacy (NAAL) Supplemental Studies." *National Center for Education Statistics,* 2009.

[6] R. C. Anderson *et al*, "Becoming a Nation of Readers: The Report of the Commission on Reading," *Education and Treatment of Children,* vol. 11, *(4),* pp. 389-396, 1988.

[7] C. E. Snow, S. M. Burns and P. Griffin, *Preventing Reading Difficulties in Young Children.* Washington, D.C.: The National Academies Press, 1998.

[8] M. Balmuth, *Essential Characteristics of Effective Adult Literacy Programs: A Review and Analysis of the Research.* Brooklyn, NY: Kingsborough Community College, 1986.

[9] H. Abadzi, *What We Know about Acquisition of Adult Literacy: Is There Hope?* Washington, D.C.: World Bank Group, 1994*245*.

[10] C. Perfetti and M. Marron, "Learning to read: literacy acquisition by children and adults," 1998.

[11] C. Read and L. Ruyter, "Reading and Spelling Skills in Adults of Low Literacy," *Remedial and Special Education,* vol. 6, *(6),* pp. 43-52, 1985.

[12] L. C. Bell and C. A. Perfetti, "Reading Skill," *Journal of Educational Psychology,* vol. 86, *(2),* pp. 244-255, 1994.

[13] J. Kruidenier, "Literacy Education in Adult Basic Education," vol. 3, *(2002),* pp. 1-57, 2002.

[14] M. J. Snowling and C. Hulme, *The Science of Reading: A Handbook.* 2008.

[15] M. R. Kuhn and S. A. Stahl, "Fluency: A review of developmental and remedial practices," *Journal of Educational Psychology,* vol. 95, *(1),* pp. 3-21, 2003.

[16] M. A. Curtis and J. R. Kruidenier, "Teaching Adults to Read," 2005.

[17] D. J. Chard and S. V. Dickson, "Phonological Awareness: Instructional and Assessment Guidelines," *Intervention in School and Clinic,* vol. 34, *(5),* pp. 261-270, 1999.

[18] P. A. Popp, *Reading on the Go! Students Who are Highly Mobile and Reading Instruction.* Greensboro, NC: National Center for Homeless Education at SERVE, 2004*1*.

[19] R. H. Yopp and H. K. Yopp, "Supporting Phonemic Awareness Development in the Classroom," *Reading Teacher,* vol. 54, *(2),* pp. 130-43, 2000.

[20] I. Y. Liberman *et al*, "Explicit Syllable and Phoneme Segmentation in the Young Child," *Journal of Experimental Child Psychology,* vol. 18, *(2),* pp. 201-212, 1974.

[21] V. E. Snider, "A Primer on Phonemic Awareness: What It Is, Why It's Important, and How to Teach It," *School Psychology Review,* vol. 24, *(3),* pp. 443-455, 1995.

[22] P. B. Gough and C. H. Lee, "A Step Toward Early Phonemic Awareness: The Effects of Turtle Talk Training," *Psychologia,* vol. 50, *(1),* pp. 54-66, 2007.

[23] J. Kruidenier, *Research Based Principles for Adult Basic Education Reading Instruction.* 2002.

[24] I. Y. Liberman, D. Shankweiler and A. M. Liberman, "The alphabetic principle and learning to read," in *Phonology and Reading Disability: Solving the Reading Puzzle*, I. Y. Liberman and D. Shankweiler, Eds. Ann Arbor, MI: The University of Michigan Press, 1989, pp. 1-33.

[25] D. Shankweiler, "How problems of comprehension are related to difficulties in decoding," in *Phonology and Reading Disability: Solving the Reading Puzzle*, D. Shankweiler and I. Liberman, Eds. Ann Arbor, MI: The University of Michigan Press, 1989, pp. 35-68.

[26] V. A. Mann, "Longitudinal prediction and prevention of early reading difficulty," *Annals of Dyslexia,* vol. 34, *(1),* pp. 117-136, 1984.

[27] C. Juel, "Learning to read and write: A longitudinal study of 54 children from first through fourth grades," *Journal of Educational Psychology,* vol. 80, *(4),* pp. 437-447, 1988.

[28] E. W. Ball and B. A. Blachman, "Does Phoneme Awareness Training in

Kindergarten Make a Difference in Early Word Recognition and Developmental Spelling?" *Reading Research Quarterly,* vol. 26, *(1),* pp. 49-66, 1991.

[29] A. Cohen and R. Horowitz, "What should teachers know about bilingual learners and the reading process?" in *Literacy and the Second Language Learner*, J. H. Sullivan, Ed. Greenwich, CT: Information Age Publishing, 2002, pp. 29-53.

[30] A. C. Pratt and S. Brady, "Relation of Phonological Awareness to Reading Disability in Children and Adults," *Journal of Educational Psychology,* vol. 80, *(3),* pp. 319-323, Sep, 1988.

[31] J. Morais *et al*, "Does awareness of speech as a sequence of phones arise spontaneously?" *Cognition,* vol. 7, *(4),* pp. 323-331, 1979.

[32] C. A. Perfetti and M. A. Marron, "Learning to read: literacy acquisition by children and adults," *Advances in Adult Literacy Research and Development,* pp. 2-42, 1998.

[33] C. Read *et al*, "The ability to manipulate speech sounds depends on knowing alphabetic writing," *Cognition,* vol. 24, *(1),* pp. 31-44, November 1, 1986.

[34] T. W. Hogaboam and C. A. Perfetti, "Lexical ambiguity and sentence comprehension," *Journal of Verbal Learning and Verbal Behavior,* vol. 14, *(3),* pp. 265-274, June 1, 1975.

[35] A. M. Liberman *et al*, "Linguistic abilities and spelling proficiency in kindergarten and adult poor spellers," in *Biobehavioural Measures of Dyslexia*, D. Gray and J. Kavanaugh, Eds. Parkton, MD: York Press, 1985, pp. 163-176.

[36] R. Wagner, C. Schatschneider and C. Phythian-Sense, *Beyond Decoding: The Behavioral and Biological Foundations of Reading Comprehension.* 2009.

[37] D. Greenberg, L. C. Ehri and D. Perin, "Are word-reading processes the same or different in adult literacy students and third-fifth graders matched for reading level?" *J. Educ. Psychol.,* vol. 89, *(2),* pp. 262-275, 1997.

[38] D. J. Chard, J. J. Pikulski and S. Templeton, "From Phonemic Awareness to Fluency: Effective decoding instruction in a research-based reading program," *Houghton Mifflin,* pp. 1-12, 2000.

[39] D. Perin, "Phonemic segmentation and spelling," *British Journal of Psychology*, vol. 74, *(1),* pp. 129-144, February 1, 1983.

[40] National Reading Panel, "Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction," *NIH Publication no.00-4769,* vol. 7, pp. 35, 2000.

[41] J. K. Torgesen *et al*, "Intensive remedial instruction for children with severe reading disabilities: immediate and long-term outcomes from two instructional approaches," *Journal of Learning Disabilities,* vol. 34, *(1),* pp. 58, 78, Jan-Feb, 2001.

[42] D. LaBerge and S. J. Samuels, "Toward a theory of automatic information processing in reading," *Cognit. Psychol.,* vol. 6, *(2),* pp. 293-323, 1974.

[43] E. D. Hirsch, "Reading Comprehension Requires Knowledge — of Words and the World," *American Educator,* pp. 10-45, 2003.

[44] J. J. Pikulski and D. J. Chard, "Fluency: Bridge Between Decoding and Reading Comprehension," *The Reading Teacher,* vol. 58, *(6),* pp. 510-519, 2005.

[45] P. Yovanoff *et al*, "Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency," *Educational Measurement: Issues and Practice,* vol. 24, *(3),* pp. 4-12, 2005.

[46] P. N. Johnson-Laird, *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* 1983.

[47] W. Kintsch, "How readers construct situation models for stories: The role of syntactic cues and causal inferences," in *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes, Volume II,* A. F. Healy, S. M. Kosslyn and R. M. Shiffrin, Eds. Hillsdale, NJ: Psychology Press, 1992, pp. 261-278.

[48] R. A. Zwaan and G. A. Radvansky, "Situation models in language comprehension and memory," *Psychological Bulletin,* vol. 123, *(2),* pp. 162-185, Mar, 1998.

[49] S. A. Stahl and M. M. Fairbanks, "The Effects of Vocabulary Instruction: A Model-Based Meta-Analysis," *Review of Educational Research,* vol. 56, *(1),* pp. 72-110, 1986.

[50] K. Cain, "Story knowledge and comprehension skill," in *Reading Comprehension Difficulties: Processes and Intervention*, C. Cornoldi and J. Oakhill, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 1996, pp. 167-192.

[51] L. M. Hoffman, "Reading Errors among Skilled and Unskilled Adult Readers," *Community Junior College Research Quarterly,* vol. 2, *(2),* pp. 151-162, January 1, 1978.

[52] J. S. Chall, *Stages of Reading Development.* New York, NY: McGraw-Hill, 1983.

[53] K. E. Stanovich, "Matthew Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy," *Reading Research Quarterly,* vol. 22, *(1),* pp. 360-407, 1986.

[54] W. E. Nagy and J. A. Scott, *Vocabulary Processes.* Mahwah, NJ: Erlbaum, 2000.

[55] A. Newman, T. Rosbash and L. Sarkisian, "Learning for life: The opportunity for technology to transform adult education (part II)," 2015.

[56] ProLiteracy, "2015-16 Annual Statistical Report," 2016.

[57] D. D. Amstutz and V. Sheared, "The Crisis in Adult Basic Education," *Education and Urban Society,* vol. 32, *(2),* pp. 155-166, 2000.

[58] S. Kerka, "Adult Learner Retention Revisited," *ERIC Digest, (166),* pp. 1-8, 1995.

[59] M. B. Young *et al*, *National Evaluation of Adult Education Programs: Executive Summary.* Arlington, VA: Development Associates, 1995.

[60] J. Comings and L. Sorricone, "Massachusetts: A case study of improvement and growth of adult education services," in *The Review of Adult Learning and Literacy, Volume 5*, J. Comings, B. Garner and C. Smith, Eds. Mahwah, NJ: Lawrence Erlbaum, 2005, pp. 85-123.

[61] B. A. Quigley, "Understanding and overcoming resistance to adult literacy education," Institute for the Study of Adult Literacy, The Pennsylvania State University, University Park, PA, 1992.

[62] A. Newman, T. Rosbash and L. Sarkisian, "Learning for life: The opportunity for technology to transform adult education (part I)," 2015.

[63] A. Smith. The smartphone difference. Pew Research Center. 2015 Available: http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/.

[64] I. S. Kirsch *et al*, "Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey." *National Center for Education Statistics,* pp. 178, 2002.

[65] J. R. Welch and K. DiTommaso, "Youth in ABE: The Numbers," *Focus on Basics,* vol. 7, *(1),* pp. 18-21, 2004.

[66] J. Weber, "Youth Cultural Competence: A Pathway for Achieving Outcomes with Youth," *Focus on Basics,* vol. 7, *(A),* pp. 6-10, 2004.

[67] B. Garner, "Separate Yet Happy," *Focus on Basics,* vol. 7, *(A),* pp. 29-30, 2004.

[68] S. Bartlett, G. Kondrak and C. Cherry, "On the Syllabification of Phonemes," *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, (June),* pp. 308-316, 2009.

[69] W. M. Fisher, "A statistical text-to-phone function using n-grams and rules," in *1999 IEEE International Conference,* Washington, D.C., 1999, pp. 649-652.

[70] Daryl Bullis, "Oxford English Dictionary," Sep 3, 2015.

[71] D. Greenberg, L. C. Ehri and D. Perin, "Do Adult Literacy Students Make the Same Word-Reading and Spelling Errors as Children Matched for Word-Reading Age?" *Scientific Studies of Reading,* vol. 6, *(3),* pp. 221-243, Jul 1, 2002.

[72] G. Spache, "Characteristic Errors of Good and Poor Spellers," *The Journal of Educational Research,* vol. 34, *(3),* pp. 182-189, 1940.

[73] K. Sakaguchi, Y. Arase and M. Komachi, "Discriminative approach to fill-in-the-blank quiz generation for language learners," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,* 2013, pp. 238-242.

[74] J. C. Brown, G. A. Frishkoff and M. Eskenazi, "Automatic question generation for vocabulary assessment," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing,* 2005, pp. 826.

[75] J. Pino and M. Eskenazi, "Semi-Automatic Generation of Cloze Question Distractors Effect of Students' L1," *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education,* pp. 1-4, 2009.

[76] T. Goto *et al*, "Automatic generation system of multiple-choice cloze questions and its evaluation," *Knowledge Management and E-Learning,* vol. 2, *(3),* pp. 210-224, 2010.

[77] M. Agarwal and P. Mannem, "Automatic gap-fill question generation from text books," in *Proceedings of the 6th Workshop on Innovative use of NLP for Building Educational Applications,* 2011, pp. 56-64.

[78] N. Karamanis, L. A. Ha and R. Mitkov, "Generating multiple-choice test items from medical text: A pilot study," in *Proceedings of the Fourth International Natural Language Generation Conference,* 2006, pp. 111-113.

[79] T. Zesch and O. Melamud, "Automatic generation of challenging distractors using context-sensitive inference rules," in *Proceedings of the 9th Workshop on Innovative use of NLP for Building Educational Applications,* 2014, pp. 143-148.

[80] G. Kumar, R. E. Banchs and L. F. D'Haro, "RevUP: Automatic gap-fill question generation from educational texts," in *Proceedings of the 10th Workshop on Innovative use of NLP for Building Educational Applications,* 2015, pp. 154-161.

[81] I. Aldabe, M. Maritxalar and R. Mitkov, "A study on the automatic selection of candidate sentences distractors," *Frontiers in Artificial Intelligence and Applications,* vol.

200, *(1),* pp. 656-658, 2009.

[82] J. Mostow and H. Jang, "Generating diagnostic multiple choice comprehension cloze questions," in *Proceedings of the 7th Workshop on Innovative use of NLP for Building Educational Applications*, 2012, pp. 136-146.

[83] Jean-Baptiste Michel et al, "Quantitative Analysis of Culture Using Millions of Digitized Books," Science, vol. 331, (6014), pp. 176-182, 2011. Available: http://www.jstor.org/stable/40986490. DO

[84] B. Hannon and M. Daneman, "Facilitating Knowledge-Based Inferences in Less-Skilled Readers," *Contemporary Educational Psychology*, vol. 23, *(2),* pp. 149-172, April 1, 1998.

[85] K. Cain *et al*, "Comprehension skill, inference-making ability, and their relation to knowledge," *Memory & Cognition,* vol. 29, *(6),* pp. 850-859, Sep, 2001.

[86] L. Baker, "Comprehension monitoring: Identifying and coping with text confusions," *Journal of Reading Behavior*, vol. 11, *(4),* pp. 365-374, 1979.

[87] J. Oakhill, "Inferential and Memory Skills in Children's Comprehension of Stories," *British Journal of Educational Psychology*, vol. 54, *(1),* pp. 31-39, 1984.

[88] D. Coniam, "A Preliminary Inquiry Into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests," *CALICO Journal*, vol. 14, *(2)*, pp. 15-33, 1997.

[89] C. Shei, "FollowYou!: An Automatic Language Lesson Generation System," *Computer Assisted Language Learning*, vol. 14, *(2),* pp. 129-144, 2001.

[90] R. Mitkov, L. An Ha and N. Karamanis, "A computer-aided environment for generating multiple-choice test items," *Natural Language Engineering*, vol. 12, *(02),* pp. 177, 2006.

[91] T. Shanahan, M. L. Kamil and A. W. Tobin, "Cloze as a Measure of Intersentential Comprehension," *Reading Research Quarterly*, vol. 17, *(2),* pp. 229-255, 1982.

[92] J. Oakhill and N. Yuill, "Pronoun resolution in skilled and less-skilled comprehenders: Effects of memory load and inferential complexity," *Language and Speech*, vol. 29, *(1),* pp. 25-37, 1986.

[93] J. Pennington, R. Socher and C. D. Manning, "GloVe: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532-1543, 2014.

[94] K. Toutanova *et al*, "Feature-rich part-of-speech tagging with a cyclic dependency network," in 2003, pp. 252-259.

[95] K. Cain, J. Oakhill and C. Elbro, "The ability to learn new word meanings from context by school-age children with and without language comprehension difficulties," *Journal of Child Language,* vol. 30, *(3),* pp. 681-694, Aug 1, 2003.

[96] K. Cain, "Text comprehension and its relation to coherence and cohesion in children's fictional narratives," *British Journal of Developmental Psychology,* vol. 21, pp. 335-351, 2003.

[97] K. Cain, N. Patson and L. Andrews, "Age- and ability-related differences in young readers' use of conjunctions," *Journal of Child Language,* vol. 32, *(4),* pp. 877, 2005.

[98] A. C. Crosson and N. K. Lesaux, "Does knowledge of connectives play a unique role in the reading comprehension of English learners and English-only students?" *Journal of Research in Reading,* vol. 36, *(3),* pp. 241-260, 2013.

[99] Z. Lin, H. T. Ng and M. Kan, "A PDTB-styled end-to-end discourse parser," *Natural Language Engineering,* vol. 20, *(2),* pp. 151, Apr 1, 2014.

[100] R. Prasad *et al*, "The Penn Discourse Treebank 2.0," in *Language Resources and Evaluation,* Morocco, 2008.

[101] Ö Esit, "Your verbal zone: an intelligent computer-assisted language learning program in support of Turkish learners' vocabulary learning," *Computer Assisted Language Learning,* vol. 24, *(3),* pp. 211-232, July 1, 2011.

[102] B. Gorjian *et al*, "The impact of asynchronous computer-assisted language learning approaches on English as a foreign language high and low achievers' vocabulary retention and recall," *Computer Assisted Language Learning,* vol. 24, *(5),* pp. 383-391, 2011.

[103] C. C. Iheanacho, "Effects of Two Multimedia Computer-Assisted Language Learning Programs on Vocabulary Acquisition of Intermediate Level ESL Students." , Virginia Polytechnic Institute and State University, 2002.

[104] Z. Li and V. Hegelheimer, "Mobile-assisted grammar exercises: Effects on self-editing in L2 Writing," *Language Learning & Technology,* vol. 17, *(3),* pp. 135-156, 2013.

[105] P. Pirasteh, "The Effectiveness of Computer-assisted Language Learning (CALL) on Learning Grammar by Iranian EFL Learners," *Procedia - Social and Behavioral Sciences,* vol. 98, *(Supplement C),* pp. 1422-1427, May 6, 2014.

[106] A. M. Karemaker, N. J. Pitchford and C. O'Malley, "Does whole-word multimedia software support literacy acquisition?" *Reading and Writing,* vol. 23, *(1),* pp. 31-51, 2010.

[107] K. Regan *et al,* "Effects of Computer-Assisted Instruction for Struggling Elementary Readers With Disabilities," *The Journal of Special Education,* vol. 48, *(2),* pp. 106-119, 2014.

[108] S. A. Whitcomb, J. D. Bass and J. K. Luiselli, "Effects of a Computer-Based Early Reading Program (Headsprout) on Word List and Text Reading Skills in a Student with Autism," *Journal of Developmental and Physical Disabilities,* vol. 23, *(6),* pp. 491-499, 2011.

[109] J. Cassady and L. Smith, "The Impact of a Reading-Focused Integrated Learning System on Phonological Awareness in Kindergarten," *Journal of Literacy Research,* vol. 35, *(4),* pp. 947-964, 2003.

[110] S. A. Hecht and L. Close, "Emergent literacy skills and training time uniquely predict variability in responses to phonemic awareness training in disadvantaged kindergartners," *Journal of Experimental Child Psychology,* vol. 82, *(2),* pp. 93-115, 2002.

[111] P. Macaruso and A. Walker, "The efficacy of computer-assisted instruction for advancing literacy skills in kindergarten children," *Reading Psychology,* vol. 29, *(3),* pp. 266-287, 2008.

[112] E. Segers and L. T. Verhoeven, "Long-term effects of computer training of phonological awareness in kindergarten," *Journal of Computer Assisted Learning,* vol. 21, *(1),* pp. 17-27, 2005.

[113] J. V. D'Agostino *et al,* "Introducing an iPad app into literacy instruction for struggling readers: Teacher perceptions and student outcomes," *Journal of Early Childhood Literacy,* vol. 16, *(4),* pp. 522-548, 2016.

[114] E. Segers and L. Verhoeven, "Multimedia support of early literacy learning," *Computers & Education,* vol. 39, *(3),* pp. 207-221, 2002.

[115] S. McKenney and J. Voogt, "Designing technology for emergent literacy: The PictoPal initiative," *Computers & Education,* vol. 52, *(4),* pp. 719-729, 2009.

[116] E. Gu, M. H. Tzou and R. Hoda, "Fill that blank! An iOS-based literacy application," *Proceedings of the Australian Software Engineering Conference, ASWEC,* pp. 80-83, 2014.

[117] J. Horton, D. Ellis and P. Black, "the Design and Development of an Intelligent Tutoring System for Adult Literacy Students,"*Computer Assisted Language Learning,*

vol. 2, *(1),* pp. 69-81, 1990.

[118] C. Munteanu *et al*, "ALEX: Mobile language assistant for low-literacy adults," in *12th International Conference on Human Computer Interaction with Mobile Devices and Services*, 2010, pp. 427-430.

[119] W. M. Watanabe *et al*, "Facilita: Reading assistance for low-literacy readers," *Proceedings of the 27th ACM International Conference on Design of Communication - SIGDOC '09, (January),* pp. 29, 2009.

[120] K. Browne, C. Anand and E. Gosse, "Gamification and serious game approaches for adult literacy tablet software," *Entertainment Computing,* vol. 5, *(3),* pp. 135-146, 2014.

[121] C. Ksoll *et al*, "Learning without teachers? A randomized experiment of a mobile phone-based adult education program in Los Angeles," Center for Global Development, Washington, D.C., May 22. 2014.

[122] S. Ramachandran and R. Stottler, "An intelligent tutoring system for adult literacy enhancement," in *Proceedings of the Fifth International Conference on Intelligent Tutoring Systems*, Montreal, 2000, pp. 461-477.

[123] I. Medhi, A. Sagar and K. Toyama, "Text-free user interfaces for illiterate and semi-literate users," in *International Conference on Information and Communication Technology and Development,* 2006, pp. 72-82.

[124] T. Parikh, K. Ghosh and A. Chavan, "Design studies for a financial management system for micro-credit groups in rural India," University of Washington, Jan. 2002.

[125] I. Medhi *et al*, "Designing mobile interfaces for novice and low-literacy users," *ACM Transactions on Computer-Human Interaction,* vol. 18, *(1),* pp. 1-28, 2011.

[126] E. Friscira, H. Knoche and J. Huang, "Getting in touch with text: Designing a mobile phone application for illiterate users to harness SMS," in *Proceedings of the 2nd ACM Symposium on Computing for Development,* Atlanta, GA, 2012.

[127] K. Ghosh, T. Parikh and A. Chavan, "Design considerations for a financial management system for rural, semi-literate users," in Ft Lauderdale, FL, Apr 5, 2003, pp. 824-825.

[128] I. Medhi-Thies *et al*, "KrishiPustak: A social networking system for low-literate farmers," in Vancouver, BC, Canada, Feb 28, 2015, pp. 1670-1681.

[129] W. Stuifmeel, "Mobile UI Design for Low Literacy Users in West Africa." 2009.

[130] *OpenClipart*. Available: https://openclipart.org/.

[131] Van Linden Sabine and A. H. M. Cremers, "Cognitive abilities of functionally illiterate persons relevant to ICT use," *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* vol. 5105 LNCS, pp. 705-712, 2008.

[132] A. Chand, "Designing for the Indian rural population: Interaction design challenges," in Bangalore, India, 2002.

[133] M. P. Huenerfauth, "Developing Design Recommendations for Computer Interfaces Accessible to Illiterate Users.", 2002.

[134] S. Grisedale, M. Graves and A. Grünsteidl, "Designing a graphical user interface for healthcare workers in rural India," in New York, NY, 1997, pp. 471–478.

[135] Eshet-Alkalai, "Digital Literacy: A Conceptual Framework for Survival Skills in the Digital Era," *Journal of Educational Multimedia and Hypermedia,* vol. 13, *(1),* pp. 93-106, 2004.

[136] B. M. Chaudry *et al*, "Mobile interface design for low-literacy populations," in *Proceedings of the 2nd ACM SIGHIT Symposium on International Health Informatics,* 2012, pp. 91-100.

[137] K. Summers and M. Summers, "Reading and navigational strategies of web users with lower literacy skills," in *Proceedings of the American Society for Information Science and Technology,* Charlotte, 2005.

[138] M. Bigelow and R. Schwarz. Adult English language learners with limited literacy. National Institute for Literacy. Washington, D.C. 2010.

[139] D. H. Schunk, "Self-Efficacy for Reading and Writing: Influence of Modeling, Goal Setting, and Self-Evaluation," *Reading and Writing Quarterly: Overcoming Learning Difficulties,* vol. 19, *(2),* pp. 159-72, 2003.

[140] A. T. Corbett and J. R. Anderson, "Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes," in New York, NY, 2001, pp. 245–252.

[141] I. Medhi and K. Toyama, "Full-context videos for first-time, non-literate PC users," in Bangalore, India, 2007, pp. 1-9.

[142] J. Hamari, J. Koivisto and H. Sarsa, "Does gamification work? -- A literature review of empirical studies on gamification," in 2014, pp. 3025-3034.

[143] S. Yoon *et al*, "Active learning, deliberate practice, and educational technology

in professional education: Practices and implications," in *Handbook of Research on Educational Technology Integration and Active Learning*, J. Keengwe, Ed. Hershey, PA: IGI Global, 2015, pp. 177-201.

[144] B. Cambourne, "The whole story: Natural learning and the acquisition of literacy in the classroom," in Anonymous New York: Ashton Scholastic, 1988, pp. 216.

[145] K. Toukonen, "The Dynamic Electronic Textbook: Enhancing the Student's Learning Experience." Kent State University, 2011.

Appendix A - ARPAbet phoneme set

Phoneme	Example	Phonetic Breakdown
/AA/	odd	AA D
/AE/	at	AE T
/AH/	hut	HH AH T
/AO/	ought	AO T
/AW/	cow	K AW
/AY/	hide	HH AY D
/B/	be	B IY
/CH/	cheese	CH IY Z
/D/	dee	D IY
/DH/	thee	DH IY
/EH/	Ed	EH D
/ER/	hurt	HH ER T
/EY/	ate	EY T
/F/	fee	F IY
/G/	go	G OW
/HH/	he	HH IY
/IH/	it	IH T
/IY/	eat	IY T
/JH/	gee	JH IY
/K/	key	K IY
/L/	lee	L IY
/M/	me	M IY
/N/	knee	N IY
/NG/	ping	P IH NG
/OW/	oat	OW T
/OY/	toy	T OY
/P/	pee	P IY
/R/	read	R IY D
/S/	sea	S IY
/SH/	she	SH IY
/T/	tea	T IY
/TH/	they	TH EY
/UH/	hood	HH UH D
/UW/	two	T UW
/V/	vee	V IY
/W/	we	W IY
/Y/	yes	Y EH S
/Z/	zee	Z IY
/ZH/	seize	S IY ZH

Appendix B - Top three letter mappings for each vowel phoneme by frequency

/AA/	/AE/	/AH/	/AO/	/AW/	/AY/
o 67.6%	*a* 99.7%	*u* 32.8%	*o* 57.8%	*ou* 62.8%	*i* 86.2%
a 29.9%		*e* 16.0%	*a* 16.7%	*ow* 37.2%	*y* 9.8%
		a 16.9%	*au* 12.2%		

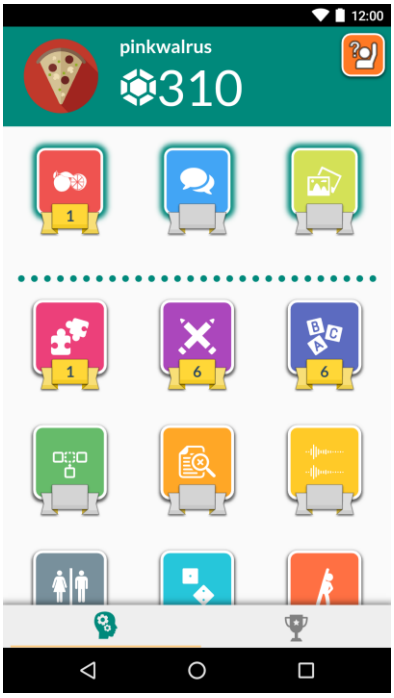
/EH/	/ER/	/EY/
e 67.6%	*er* 44.0%	*a* 75.7%
a 29.9%	*ur* 19.5%	*ai* 10%
ea 5.3%	*or* 9.7%	*ay* 8.4%

/IH/	/IY/
i 80.0%	*y* 35.2%
e 15.1%%	*e* 25.7%
	ea 15.7%

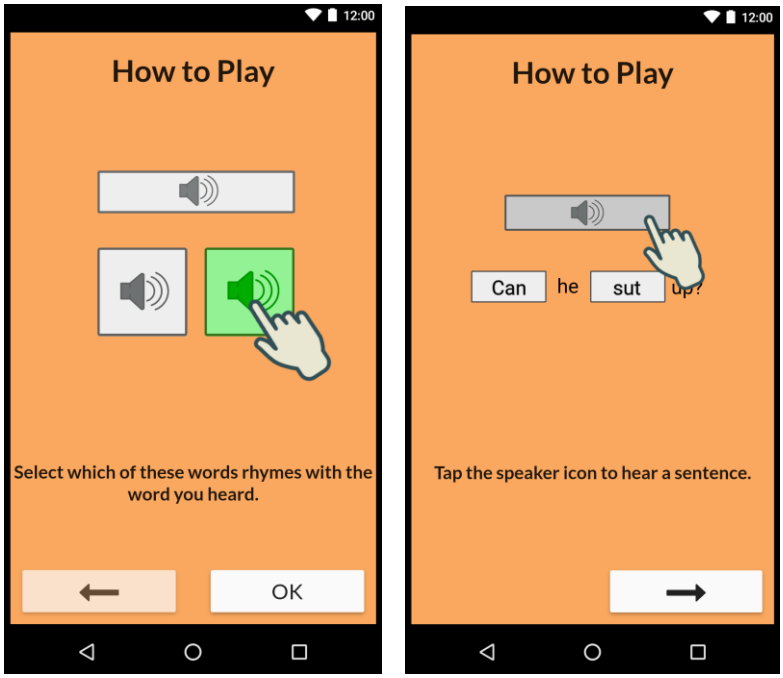
/OW/	/OY/
o 74.0%	*er* 54.0%
ow 12.6%	*ur* 46.0%
oa 9.5%	

/UW/	/UH/
u 40.6%	*oo* 87.9%
oo 27.8%	*u* 12.1%
ew 11.7%	

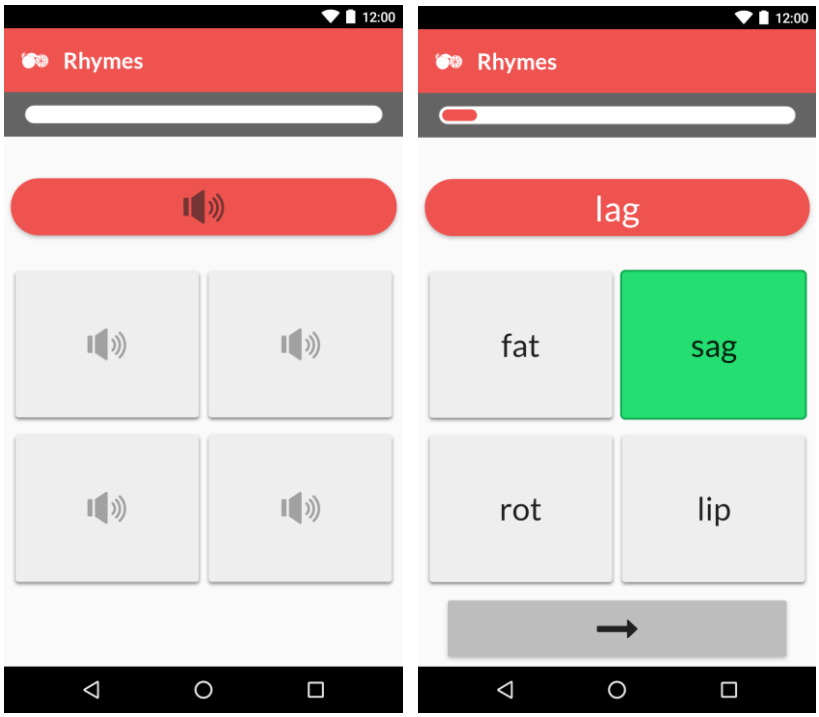
Appendix C - Screenshots of the CAPITAL app



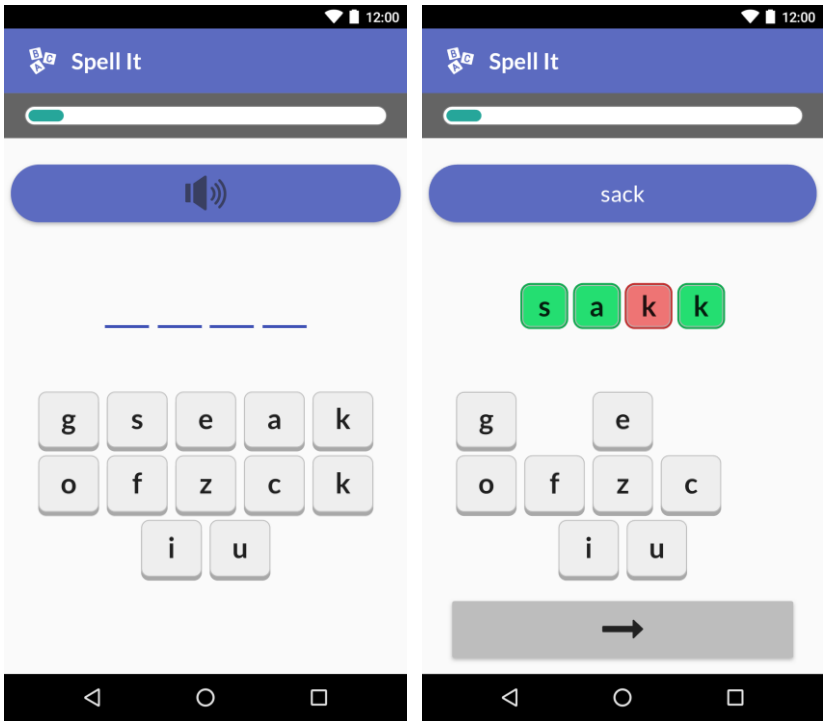
Dashboard with all exercises, and the user's levels for the exercises that they have taken placement tests for



Examples of tutorial screens that users see before every exercise, with instructions read aloud

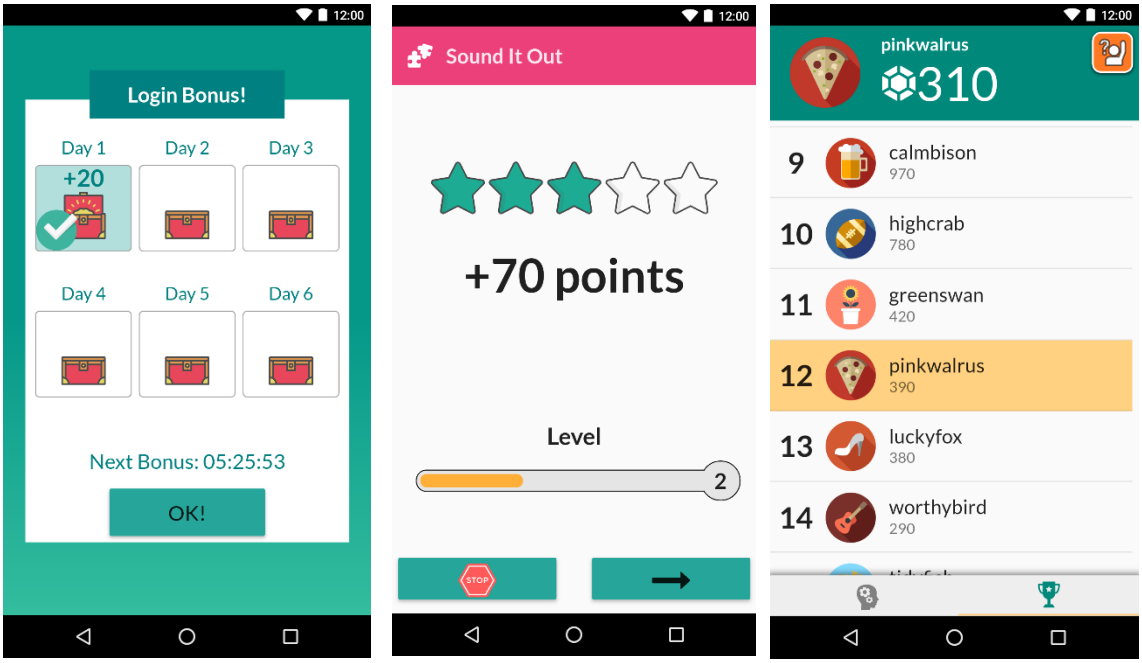


Exercise: Find the Rhyme - Choose the word that rhymes with the given word



Exercise: Spell It - Spell a word given its audio

Exercise: Pick the Misspelling - Identify a misspelled word in a sentence



Gamification: daily login bonus (left); star meter and points earned (middle); leaderboard (right)

Appendix D - Software usability survey for ABE instructors

The following survey will ask you a series of questions to get your feedback on the user interface for the CAPITAL app. The survey is broken into several sections, each of which contains a small number of opinion-based questions related to a particular aspect of the app.

Each question presents a statement about the app and asks you to rate your agreement with that statement on a 1 to 5 scale, with 1 being "Strongly disagree" and 5 being "Strongly agree".

User Experience

Emotional issues

The lessons are motivating and fun.

The app encourages user participation.

The app experience is enjoyable and exciting.

It is new technology, yet it is an interesting and acceptable form of learning.

Contextual factors

This type of learning suits the needs of the user.

Prior experience with mobile devices makes using the app easy.

The target user base has been considered in the design of the app.

Goals are clearly set and fixed.

User-centricity

The app allows for personalized learning.

Experimentation and exploration is possible.

Requirements are made clear to the user.

The user can be self-sufficient.

Students can customize their learning experience.

Users are required to use active learning and critical thinking.

Users are able to direct their own learning with a sense of ownership.

Appeal

The learning environment is stimulating.

The student is motivated to explore.

The app is visually appealing.

Satisfaction

The app makes the learning experience fun.

The user feels a sense of achievement and accomplishment when using the app.

The user is encouraged to engage with the material.

General Interface Usability

Visibility of system status

The system is responsive to user actions without odd and unexplained events.

Visible feedback clearly communicates what is happening.

Match to the real world

The application uses language that is appropriate for the target users.

Information is presented in a sequential, logical, and naturally arranged way.

Learner control and freedom

It is possible to exit the app at any time.

Consistency; adherence to standards

Symbols and icons are used logically and consistently throughout the app.

Prevention of usability-related errors

The system is designed to prevent errors from occurring.

Clear and appropriate feedback is shown if a mistake is made.

Recognition rather than recall

Minimal scrolling is needed to see all elements on the screen.

Aesthetics and minimalism in design

The app does not include distracting or irrelevant material.

Graphics and icons are used to illustrate a point rather than to decorate the page.

Recognition and recovery from errors

Error messages are easy to follow and presented with clear language.

Quick and simple solutions are offered if errors are made.

The user can easily recover from errors as a result of the feedback given by the app.

Help and documentation

The help feature is easy to find and supports the user's needs.

Help/support is provided on each page.

Web-Based Learning

Simple, well-organized navigational structure

The application is easy to navigate on a mobile device.

There are several paths to and from a chosen destination.

Related information has been grouped into obvious categories.

Information is organized hierarchically.

Links and buttons support navigation throughout the site without cluttering it.

Relevant pedagogical site content

The app is interesting and keeps the user's attention focused.

The information in the app is clear and relevant.

No racial or gender biases are noted.

Suitable course content of a high quality

The content provided is of a high standard.

Easy-to-use system

No difficulties are experienced in trying to reach materials.

It is easy to navigate back and forth through the learning options offered.

Educational Usability

Clarity of goals, objectives, and outcomes

Goals are clearly set out, and objectives and expected outcomes for learning are clear.

There is a good reason for the inclusion of each page and this reason is obvious.

Error recognition, diagnosis, and recovery

Users are always given the opportunity to see their mistakes and learn from them.

Feedback, guidance, and assessment

Users receive prompt feedback from the application on assessment and progress.

Guidance is provided about each learning task.

Activities are graded, with grades providing instant feedback and correction.

---

m-Learning Features

Handheld devices and technology

The mobile interface does not hamper working with the application.

The mobile platform is used to its fullest capability.

Flexibility

The lessons may be done at any time.

Lessons may be viewed in any order.

The app can be used anytime and anywhere.

Interactivity

Multiple kinds of exercises have been provided.

Interaction happens in various ways.

Interactions with the application are smooth.

Interactivity has been encouraged in creative ways.