8-2020

Developing and Validating Stealth Assessments for an Educational Game to Assess Young Dual Language Immersion Learners,Äô Reading Comprehension

Frederick J. Poole
Utah State University

DEVELOPING AND VALIDATING STEALTH ASSESSMENTS FOR AN EDUCATIONAL GAME TO ASSESS YOUNG DUAL LANGUAGE IMMERSION LEARNERS' READING COMPREHENSION

by

Frederick J. Poole

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Instructional Technology & Learning Sciences

Approved:

Jody Clarke-Midura, Ed.D. Major Professor	Joshua J. Thoms, Ph.D. Committee Member
Andrew Walker, Ph.D. Committee Member	David F. Feldon, Ph.D. Committee Member
Victor R. Lee, Ph.D. Committee Member	Janis L. Boettinger, Ph.D. Acting Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2020

ABSTRACT

Developing and Evaluating Stealth Assessments for an Educational Game to Assess Young Chinese Dual Language Immersion Learners' Reading Comprehension

by

Frederick Poole, Doctor of Philosophy

Utah State University, 2020

Major Professor: Jody Clarke-Midura, Ed.D.
Department: Instructional Technology and Learning Sciences

The purpose of this multiple paper dissertation is to design a digital game and stealth assessments within the game to assess young second language learners' Chinese reading proficiency. In Chapter 2 (Paper 1) I describe the game designed for this dissertation and how it was implemented in a dual language immersion classroom. This study found that the digital game and in-class implementation led to significant vocabulary and reading comprehension gains. Further, seven types of support that students needed while playing the game were identified. In Chapter 3 (Paper 2), I describe how educational data mining approaches, and more specifically how data-driven explorations, can provide insight into how players interact with the game and further how those interactions relate to second language proficiency and learning. In this study, I identify time on task and use of in-game tools as important indicators for learning. In addition, four subgroups of students were identified based on their gameplay styles. Finally, in Chapter 4 (Paper 3), I describe how stealth assessments were designed and

validated within the game. This study found that the stealth assessments were significantly correlated with two external measures of reading comprehension.

(232 pages)

PUBLIC ABSTRACT

Developing and Evaluating Stealth Assessments for an Educational Game to Assess Young Chinese Dual Language Immersion Learners' Reading Comprehension

Frederick Poole

The purpose of this multiple-paper dissertation is to design a digital game and stealth assessments within the game to assess young second language learners' Chinese reading proficiency. In Chapter 2 (Paper 1), I describe the game designed for this dissertation and how it was implemented in a dual language immersion classroom. This study found that the digital game and in-class implementation led to significant vocabulary and reading comprehension gains. Further, seven types of support that students needed while playing the game were identified. In Chapter 3 (Paper 2), I describe how educational data mining approaches, and more specifically, how data-driven explorations, can provide insight into how players interact with the game and further how those interactions relate to proficiency and learning. In this study, I identify time on task and use of an in-game glossing tool as important indicators for learning. In addition, four subgroups of students were identified based on their gameplay styles. Finally, in Chapter 4 (Paper 3), I describe how stealth assessments were designed and validated within the game. This study found that the stealth assessments were significantly correlated with two external measures of reading comprehension.

ACKNOWLEDGMENTS

As the saying goes, “It takes a village to raise a child.” The same could be said for cultivating and supporting a doctoral student. In my 5-year journey through the Ph.D. program at Utah State University, I have received guidance and support from many wonderful people. I would like to first thank Dr. Jody Clarke-Midura for both her support and encouragement as I explored my interests and developed scholarly skills. Without her mentorship and support I would not be where I am today.

I would also like to thank my committee members: Dr. Andy Walker, Dr. Joshua Thoms, Dr. David Feldon, and Dr. Victor Lee for their insight, support, expertise, and time as they guided me through the dissertation process. I owe much gratitude to Dr. Joshua Thoms who not only encouraged me to pursue a PhD, but also provided me with several research opportunities and support before and during my doctoral studies.

I want to also thank my ITLS peers who lent an ear when I needed to vent, a verbal lashing when I stepped out of bounds, and friendship during hard times. First, I want to thank Katarina Pantic, who was not only a valued friend but was also the first person I went to when I needed advice on how to navigate the Ph.D. obstacles; Vincent Sun for constantly challenging my ideas (even when it was tedious); and Joana Franco for her calming presence in an otherwise chaotic environment. I also want to thank Kris Borecki for his technical support through my dissertation, but more importantly for his friendship and always open door.

Finally, I want to thank Elva Li, my wife, for not only supporting and encouraging me through the Ph.D. program, but for also putting up with neurotic late-

night ramblings, insecurities, and the emotional ups and downs that are associated with completing a dissertation.

Frederick Poole

CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION TO MULTIPLE PAPER DISSERTATION

Integrating games into foreign/second language (L2) instruction is by no means novel (Baltra, 1990; Hubbard, 1991). Digital games have been shown to promote L2 learning (e.g., Ansteeg, 2015; Bytheway, 2014; Coleman, 2002; Palaiogiannis, 2014; Suh et al., 2010; Yudintseva, 2015), afford opportunities for meaningful L2 interactions (Dalton & Devitt, 2016; Peterson, 2011; Warschauer & Healey, 1998), and provide immediate feedback in context (Cornillie et al., 2012). Over the last 20 years, research on digital games in L2 contexts has continued to grow (Cornillie et al., 2012; Hung et al, 2018; Poole & Clarke-Midura, 2020). Despite these increases, there are a few areas of research involving digital games in L2 contexts that have yet to be explored. First, most research on digital games for L2 learning occur outside of the classroom with a focus on illustrating how the digital games can function as stand-a-lone learning tools (e.g., Calvo-Ferrer, 2017). This has led some researchers to call for more research on how digital games can be integrated into the L2 classroom (e.g., deHaan, 2019; Jones, 2020). Second, one of the advantages of using a digital game in a learning environment is the ability to collect and track player data via log files. A recent systematic review noted that only 28.5% of the studies reviewed took advantages of log data in their research (Poole & Clarke-Midura, 2020). Of those few studies, most of them used the log data to export chat logs or answers to in-game quizzes (e.g., Erhel & Jamet, 2016; Peterson, 2011; Rama et al., 2012). There is great potential for using log data in L2 game research. Such data could be used to explore how learners interact with a game and how those interactions are

associated with learning. Further, in other educational subject-content areas (e.g., mathematics, physics, science) researchers have used in-game actions to develop game-based assessments (e.g., Gibson & Clarke-Midura, 2015; Gobert et al., 2012). While L2 researchers have used in-game actions as predictors for learning (e.g., Collentine, 2011; Cornillie et al., 2012; Rankin et al., 2006), to date no study has explored the use of a digital game as an L2 assessment tool.

In this multi-paper dissertation, I applied game-based assessment approaches being implemented in other educational settings to an L2 learning context. To do this, I designed and built a digital game for students in a Chinese dual language immersion (DLI) classroom to promote vocabulary learning and reading development. I implemented the game in two DLI classrooms. As students played the game, data related to the use of vocabulary support, reading in-game texts, and gameplay behaviors were collected. I applied educational data mining (EDM) techniques to analyze these data and explored potential in-game actions and behavioral patterns that are associated with L2 reading comprehension and vocabulary growth. Finally, I constructed and validated a Bayesian Belief Network to automatically assess young Chinese DLI learners' reading comprehension proficiency.

In the following sections, I first describe the research context that sets up the need for game-based assessments for Chinese L2 learners in a DLI program. I then discuss the goals and objectives for the present study. Finally, I provide an outline for this multiple-paper dissertation.

Context

Learners in the Utah DLI program are expected to not only become proficient speakers, but also to develop literacy skills (Christian, 2011; Fortune & Tedick, 2008). These literacy skills are particularly important given that when these DLI learners reach middle and high school they are expected to learn content via the target language. Researchers have argued that by developing oral skills before learning to read, learners can rely on their oral abilities to support their understanding of a written text (Dew, 1994; Koda, 2007). This is largely based on the assumption that reading is the result of matching spoken units to a writing system (Shu & Anderson, 1999). This also assumes that the written language provides phonetic indicators to allow for such matching. These assumptions are met for DLI languages that employ an alphabetic script (e.g., Spanish, French). However, for Chinese learners, matching the oral language to the logographic script is difficult (Everson, 1998). Although the Chinese script does provide some phonetic and semantic hints, they are not always reliable and their connections to meaning or sound are not always clear.

Learning Vocabulary and Reading in Chinese

Chinese words are typically formed by a combination of two characters. While a single character, and in some cases three or four characters can represent a word, two-character words are the most common. Characters themselves generally represent ideas. For example, 电 (dian) is the character for electric and is used in words like *computer*, *elevator*, and *television;* whereas, 话 (hua) is the character for speech or speaking and is

used in words like *sentences* and *conversation*. When combining the 电 and 话 together, you get the word “electric speech” or more accurately *phone*.

Learning Chinese characters can be difficult because unlike languages with an alphabetic script, there is not a direct link between the orthographic information provided in characters and the phonetic and/or semantic information associated with the character (Feldman & Siok, 1999). This is problematic, because as mentioned earlier, both L1 and L2 readers rely on their oral language skills to facilitate reading (Dew, 1994; Koda, 2007). In other words, for readers of a language with an alphabetic script, once the alphabet is learned the reader can “sound out” words and then acquire or guess their meaning by matching it to their knowledge of the oral language. However, in Chinese, one cannot simply “sound out” a character.

A small portion of Chinese characters are referred to as either pictograms or ideograms. Pictograms refer to characters that resemble an image. For instance, the character for shop or store, 店 (dian) is said to resemble a person selling something in a kiosk. Ideograms are characters that represent an idea within the character. For example, 林 (lin), the character for *forest*, is comprised of two tree radicals, and thus alone 木 (mu) indicates *wood* or a *tree* but together (林) it means a *forest*. While pictograms and ideograms can provide semantic knowledge via their orthographic representation, it is important to note that nearly 90% of all characters are phonetic compounds. These are characters that contain two parts, or radicals: a phonetic radical and a semantic radical (Wang et al., 1986). Native speakers often use phonetic and semantic radicals to either guess unknown words or as clues to known words, and learners with strong radical

knowledge have also been shown to be better readers (Hayes, 1988; Shen & Ke, 2007). It should be noted that while phonetic and semantic radicals do provide information relating to the sound and meaning of a character, this relationship is not always transparent (Feldman & Siok, 1999). For example, the silk radical 纟 found on the left side of the character 给 (gei3, give) occupies the position usually reserved for the semantic radical, but it has no apparent relation to the meaning of 给, which is "to give." Phonetic radicals are also unreliable as it has been estimated that they provide accurate clues only 26% of the time (Fan et al., 1984). As a result, many young learners and novice readers of Chinese tend to rely on simply memorizing characters.

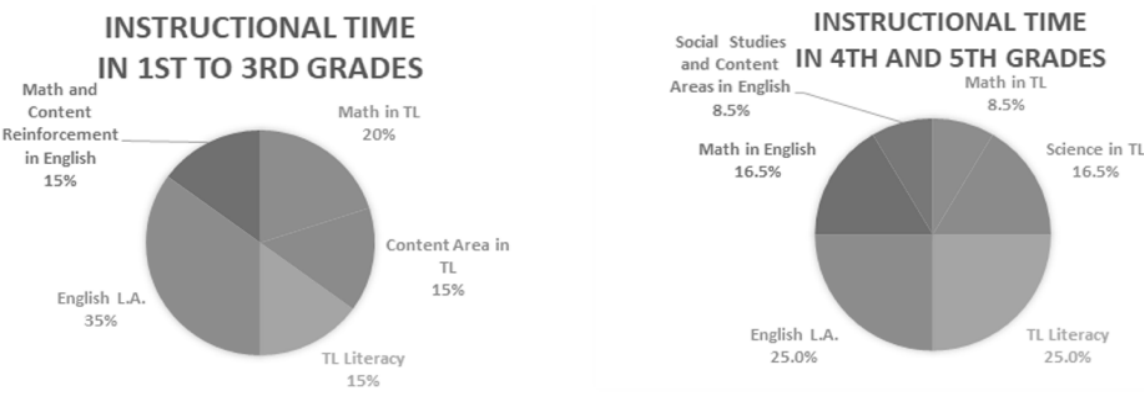
Another challenge for L2 readers of Chinese is parsing characters within sentences to form words (Shen & Jiang, 2013). Unlike languages with alphabetic scripts, there are no spaces between words. This task is further complicated because Chinese words can be comprised of one, two and/or three or more characters. Learning how to segment characters depends on story context and grammar structures. For example, the phrase 很难吃 (hen/nan/chi—very, difficult, eat), could be parsed as 很/难吃 (very disgusting) or 很/难/吃 (very difficult to eat). The differences of how to parse this phrase will be dependent on the context.

The Foreign Service Institute ranks Chinese as one of the most difficult languages for native English speakers to learn (U.S. Department of State, 2020). Unsurprisingly, two studies investigating student proficiencies in large-scale DLI programs found that Chinese DLI students' reading abilities lag behind their peers who study other languages (Burkhauser et al., 2016; Watzinger-Tharp et al., 2018). These studies have identified the

difficult script as a potential reason. Due to time constraints, DLI teachers are not able to spend more time on learning characters and developing literacy skills. Currently, a majority of their time is spent teaching subject content in the target language (e.g., math, science, social studies), see Figure 1.1. The remaining time that they do spend specifically on learning the target language must be split between both oral and literacy skills. This time constraint is further restricted when one takes into account the amount of time teachers spend on assessing their learners in the classroom. One way to address all of these issues is through digital games that provide opportunities to both practice reading and oral skills while simultaneously assessing learner proficiency.

Figure 1.1

Break down of Instructional Time



Multiple Dissertation Approach

The primary goal of this dissertation is to design and develop game-based assessments that are able to accurately evaluate young Chinese DLI learners' reading skills. To accomplish this goal, it is necessary to design a game that can be integrated into

the classroom. Further, this game must also elicit appropriate data to inform the aforementioned assessments. Finally, to continue research in this area it is important to explore other potential in-game indicators for learners and L2 proficiency growth. Thus, this dissertation is guided by the following objectives.

1. To design and evaluate a digital game used in a DLI classroom.
2. To explore how in-game actions and play styles are associated with L2 learning.
3. To develop and validate assessments embedded in the digital game.

Dissertation Outline

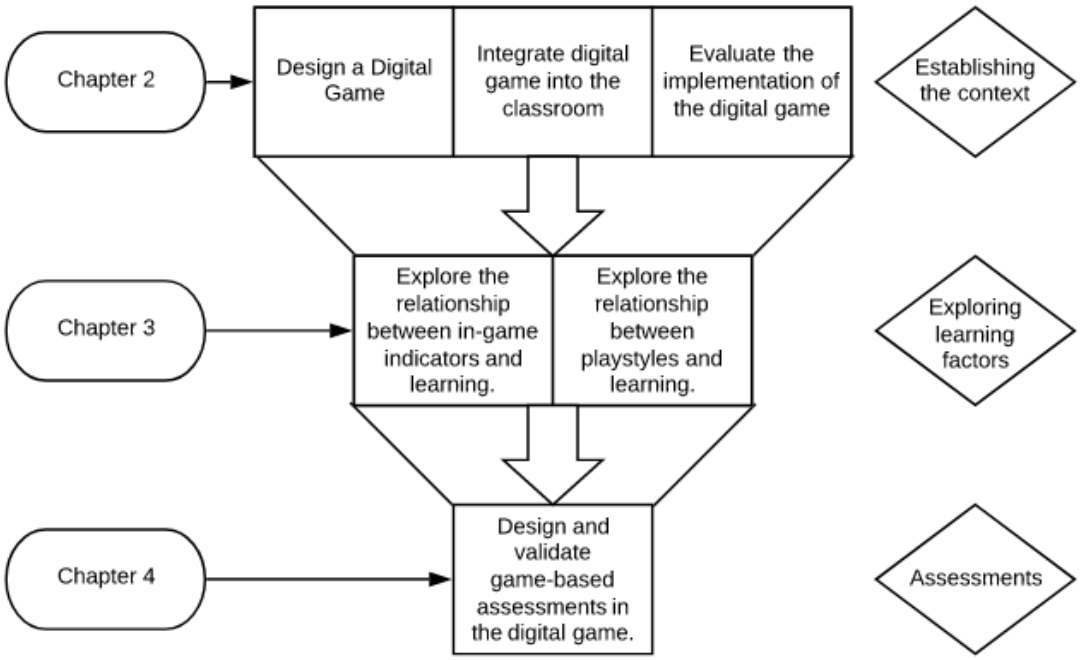
I conducted one research study yet these objectives are addressed in three separate papers. All three papers use the same participants, but focus on different data. A multiple dissertation approach was taken to contextualize each objective in its own literature and theoretical framing. While there is a clear relationship between all three of these objectives, the literature and theoretical underpinnings of each objective are different. Figure 1.2 provides an overview of the three papers in the next three chapters.

In the first paper, Chapter 2, the focus is on game design and implementation, and how they promote learning. In this first paper, I argue that past research around L2 learning and digital games has been skewed towards design and lab-like settings, without a clear focus on how such digital games can be used in L2 classrooms. In addition, I provide a review of digital games designed for L2 learning and discuss how educators have supported L2 learning in the classroom via digital games. Then, I provide detailed descriptions of the game designed for this dissertation, the supplemental material created

to support both gameplay and L2 learning, and the in-class implementation. The results of this study focus on gains in vocabulary and reading comprehension as well as the type of support that was provided by the researcher while students played the game.

Figure 1.2

Dissertation Outline



The second paper, Chapter 3, explores how educational data mining approaches, and, more specifically, how data-driven explorations can provide insight into how players interact with the game. It further explores how those interactions relate to L2 proficiency and learning. In this second paper, I review how learning analytics and educational data mining approaches have been used in L2 learning research. I then argue for the utility of

using educational data mining approaches for exploring potential relationships between gameplay behaviors and L2 proficiency and/or learning.

In the third paper, Chapter 4, drawing from the evidence-centered design framework, I discuss the design of my game-based assessment. This final paper builds off the first two papers, but focuses primarily on the design and theory that informs the assessment I developed and evaluated. I provide a detailed description of the process used to create and evaluate the assessments embedded in gameplay. Then I discuss the implications that these assessments have for educators and L2 classrooms.

Finally, in Chapter 5, I provide a summary of the findings in this dissertation and illustrate how this dissertation has addressed three currently under researched areas involving digital games for L2 learning.

CHAPTER 2

DESIGNING AND EVALUATING A DIGITAL GAME FOR THE DUAL LANGUAGE IMMERSION CLASSROOM

Abstract

This study focuses on game design and implementation and how integrating the game into a dual language immersion classroom promoted learning. I provide in-depth description of the digital game designed for this study, the supplementary material designed to support in-game learning, and the role of the teacher during in-class gameplay. Results indicate that after 4 weeks of gameplay, students had significant vocabulary and reading comprehension gains. I also coded audio data of students interacting with the instructor-researcher and identified seven types of support. This paper provides a model for one way to introduce a game into an elementary classroom setting and how to support students while they play the game.

Introduction

While many researchers have been studying the use of digital games in second language (L2) learning as stand-alone learning tools (e.g., Ansteeg, 2015; Calvo-Ferrer, 2017; Cobb & Horst, 2011; Collentine, 2011; Fotouhi-Ghazvini et al., 2009; Müller, 2012), there is a need for research on how to design and integrate games into L2 classroom instruction (e.g., deHaan, 2019; Jones, 2020). To date, there have been two distinct strands of research on digital games for L2 language and teaching: (1) game-

based, and (2) game-enhanced (Reinhardt, 2019; Reinhardt & Sykes, 2014). Game-based L2 learning and teaching refers to the use and study of games made for educational use and has focused on designing and evaluating games for L2 learning (e.g., Alyaz & Genc, 2016; Ansteeg, 2015; Cobb & Horst, 2011; Collentine, 2011; Fotouhi-Ghazvini et al., 2009; Müller, 2012). Typically, games designed in these studies target vocabulary learning (e.g., Cobb & Horst, 2011; Fotouhi-Ghazvini et al., 2009; Hung et al., 2015; McGraw et al., 2009; Müller, 2012) and there is an emphasis on illustrating how learning is promoted by the game (e.g., Calvo-Ferrer, 2017). In these studies, the game is used as an instructor and there is little emphasis or mention of the role of the teacher. Game-enhanced learning and teaching is defined as the use of vernacular or non-educational games for L2 learning and teaching. These studies have investigated how learning occurs in the wild (e.g., outside of the classroom without pedagogical support), (e.g., Piirainen-Marsh & Tainio, 2009; Scholz, 2017; Thorne et al., 2012), affordances of games in lab-like settings (e.g., Bytheway, 2014; Rankin et al., 2006; Vandercruysse et al., 2013) and in some cases how educators have used such games to promote learning in classroom settings (e.g., Miller & Hegelheimer, 2006; Ranalli, 2008; Reinders & Wattana, 2014). Again, much of this research has been positive, illustrating the many benefits that games afford learners, how unique game designs promote learning opportunities, and how integrating games into the classroom promotes learning and motivation. However, it is important to note that vernacular games are generally only appropriate for advanced L2 learners since in-game texts are written for native-speaking players. Researchers have suggested L2 learners cannot be expected to develop their language knowledge and skills

via games effectively without some form of support, especially L2 learners at lower proficiency levels (Cornillie et al., 2012; Peterson, 2011; Rankin et al., 2006).

While these two research trajectories provide a clear distinction between educational and vernacular games, the role of the instructor when using games to teach is less clear. Further, most of the research on L2 learning and games focuses on university-level students in either an English as a Second Language (ESL) or English as a Foreign Language (EFL) context with a dearth of research involving younger learners and other languages, specifically Chinese (Poole & Clarke-Midura, 2020). To address this need, in the present study I designed a digital game and a supplementary workbook to promote Chinese reading comprehension and vocabulary learning in sixth-grade classrooms. The present study takes the position that digital games, both educational and vernacular, used for educational purposes should be evaluated based on in-classroom implementations. This means that in addition to describing game design, researchers should also take into account and detail extra support that may facilitate in-class enactments. Thus, the present study explores the effectiveness of a digital game with support via supplementary material and teacher mediation as a means to promote Chinese vocabulary learning and reading comprehension in an elementary dual language immersion (DLI) classroom.

In the sections that follow, I first provide a review of the literature on L2 digital games including past research on educational games, games that were used in a classroom setting, and the supports that were used during enactment. I then introduce the game I designed for the present study, *Legend of the Dragon*. Next, I introduce my research design including my data analysis and my results. Finally, I discuss the findings

and limitations of my study.

Literature Review

Digital Game-based Language Learning

Early research on digital games used in L2 learning contexts focused on the potential learning benefits, challenges, and opportunities of games for L2 learning (Ang & Zaphiris, 2006; Baltra, 1990; Hubbard, 1991; Rankin et al., 2006). Recently, there has been a shift in L2 studies involving digital games to investigate a wider range of areas including student perspectives, change in affect, vocabulary development, and other more specific L2 proficiencies including listening, speaking, reading, and grammar. The game designed in the present study specifically targets vocabulary acquisition and reading comprehension, and thus this literature review will focus on digital games that have targeted vocabulary knowledge and reading comprehension skills.

Vocabulary Learning

Of the studies using educational games to develop L2 vocabulary knowledge, one study compared learning gains by participants who were exposed to vocabulary both in a game and outside of a game either via flashcards or a textbook (Müller, 2012). Müller reported that vocabulary learned in the game was recalled faster and more accurately than vocabulary learned outside of the game. This game relied on a drill-and-kill type of mechanic (Egenfeldt-Nielsen, 2007) in which students were given a patient with symptoms and then needed to pick the correct medication based on the English name. Feedback was given for incorrect solutions. The goal of this game was to promote

automaticity of the English words through repetition.

Two studies also compared vocabulary gains between different conditions within educational games (Franciosi et al., 2016; Peng et al., 2016). Franciosi et al. compared vocabulary gains among participants who used a flashcard system (Quizlet) while playing a game and those who did not. They found that those who used Quizlet reported better vocabulary gains. This finding suggests that vocabulary learning is improved when students have an opportunity to engage with the vocabulary in multiple contexts. Similarly, this study used a workbook to go along with the game. Peng et al. compared vocabulary gains among different group orientations, specifically competitive, cooperative, and conjunctive. The authors define conjunctive groups as those whose results are determined by the lowest performing member of the group. They found that for low-achieving students, the conjunctive group orientation lead to the highest gains in vocabulary.

Finally, two studies examined in-game vocabulary development without a comparison group or condition (Alyaz & Genc, 2016; Dourda et al., 2014). Alyaz and Genc found that participants scored significantly higher on a post-vocabulary test after playing an educational game designed to teach German for 8 weeks. The authors mention that there was additional instructional material but did not report on how that material related to learning. Further, while the participants played, they were asked to keep a journal to write down new vocabulary words, pragmatic phrases, and thoughts about the game. However, data from the journal was not explored. In another study, looking at vocabulary usage, Dourda et al. found that 45% of the words used in a reflection journal

after playing a digital game were previously unknown vocabulary words. It's important to note that both of these games were dialogue-driven, meaning players advance through the game by reading text and making decisions about the text. Thus, much of the vocabulary learning likely occurred as the result of reading in-game texts and then reflecting on the text via journals.

In summary, research on educational games has shown that vocabulary learning occurs as a result of playing digital games. Further, research shows that adding additional support while playing the game can further promote vocabulary learning. While the L2 gaming field is growing rapidly, there are a few areas that are still under researched. First, a majority of the studies mentioned above occur in either an ESL or EFL context with research in other languages (e.g., Chinese) lacking. Second, many of the studies above have focused on comparing a digital game intervention to a traditional format rather than exploring how the game is used in the classroom. When studies have implemented the use of additional supports while in a classroom setting, the researchers explored how providing a dictionary or an additional activity promoted learning rather than what the teacher did.

L2 Reading

In terms of L2 reading skills in educational digital games, only three studies explored how digital games promoted reading skills. Poole et al. (2018) explored an interactive fiction game that was designed to promote reading skills, in particular self-regulation skills, while reading. The authors found that students engaged in more metacognitive activity when they were prompted with a question in the game. Suh et al.

(2010) conducted a study using an educational Massive Multiplayer Online Role Playing Game (MMORPG) called *Nori School*. The authors compared L2 proficiency gains between a control group that received traditional instruction in the classroom, and a treatment group that learned via the MMORPG in the school computer lab. The authors found that after two months of two 40-minute sessions per week, the treatment group scored significantly higher on a post-reading assessment than the control group. It was inferred that reading gains were the result of reading the in-game dialogue. In a study that examined reading strategies by students playing a digital game in a content- and language-integrated learning classroom, Dourda et al. (2014) found that students enjoyed the game, received opportunities to develop vocabulary words, and used several different reading strategies including skimming/scanning, translating and transferring, repeating, use of imagery, and association of information and concepts. This research has found that learning via gaming environments is better than learning via non-gaming environments and that reading via a digital game can promote reading comprehension skills.

L2 Support in Gameplay

Although previous researchers have called for additional language support for L2 learners, only a few studies have investigated the effect of adding L2 learning assistance to gameplay. One study found that ESL learners who completed in-game tasks directed at vocabulary learning while playing *The Sims* performed better on post-vocabulary tests when they received language instruction before playing the game (Miller & Hegelheimer, 2006). This study illustrates the benefit of adding L2 pedagogical support to a digital game. Hitosugi, Schmidt, and Hayashi (2014) investigated how a vernacular game called

Food Force facilitated vocabulary development for Japanese as a foreign language learners. Similar to the Miller and Hegelheimer study, Hitosugi et al. compared vocabulary gains between a group who received support (vocabulary lists and pre-reading prompts) and a group who received no additional support. Finally, Dourda et al.'s (2014) study was one of the few projects that investigated the use of a game in a classroom setting with the support via in-class instruction prior to gameplay and post-game play activities that asked students to speak and write about their gameplay. The authors found that students improved English vocabulary knowledge, reading and writing skills, as well as subject content knowledge related to geography.

Past research has shown the benefits that a digital game can have on vocabulary learning and reading comprehension. Further, research suggests that adding pedagogical support, either via reviewing vocabulary, discussing gameplay, or recording gameplay in a diary, can be beneficial to the in-class L2 learning experience. However, research involving the type of support that teachers can provide is lacking. The present study set out to design a digital game and supplementary workbook to teach Chinese reading comprehension and vocabulary and then explore its integration into a classroom context.

Theoretical Framework

The present study adopts a sociocultural framework for learning, more specifically, learning is understood as occurring via the learners' activity with the L2 through mediation that is culturally constructed. Student learning in the present study is argued to be mediated by the digital game, the workbook, and the instructors. According

to Vygotsky's Sociocultural theory (SCT), human activity is mediated by culturally constructed artifacts and concepts (Vygotsky, 1978). In this context, mediation is defined as "the creation and use of artificial auxiliary means of acting – physically, socially, and mentally" (Lantolf, 2011, p. 25). In a classroom, a learner's activity is mediated by a teacher who formulates questions and activities in certain ways (often culturally derived), by textbooks, which similarly contain culturally constructed support and presentation structures, and by classmates among others potential sources. While there are several potential cultural artifacts that may affect what and how a student learns, the principal idea is that students do not learn and produce language in a vacuum. Whether in the classroom, in the home, or on the street, an L2 speaker will be provided with some form of mediation. Thus, in the present study I provide detailed descriptions of the game, the workbook, and the support provided by the instructor-researcher while students played the game in order to contextualize and detail how such mediation provides opportunities for learning.

When bringing games into classrooms, it is important to not only research whether or not learning happened but to also ensure that students who don't have any prior gaming experience learn just as much as students who have prior gaming experience. In other words, it is not only important to consider the classroom context, but also the knowledge and experiences that the student brings to the classroom. Thus, this research is guided by three research questions. The first two questions focus on student learning and whether or not prior gaming experience affected learning gains. The third question is focused on the types of support students need as they play the game in their

classroom.

1. Do students who play the digital game show learning gains in vocabulary or reading comprehension, as measured by the pre-and-post assessments?
2. Do students who have prior gaming experience see greater gains in vocabulary or reading comprehension?
3. What supports, if any, do students need from instructors when integrating the digital game into the classroom?

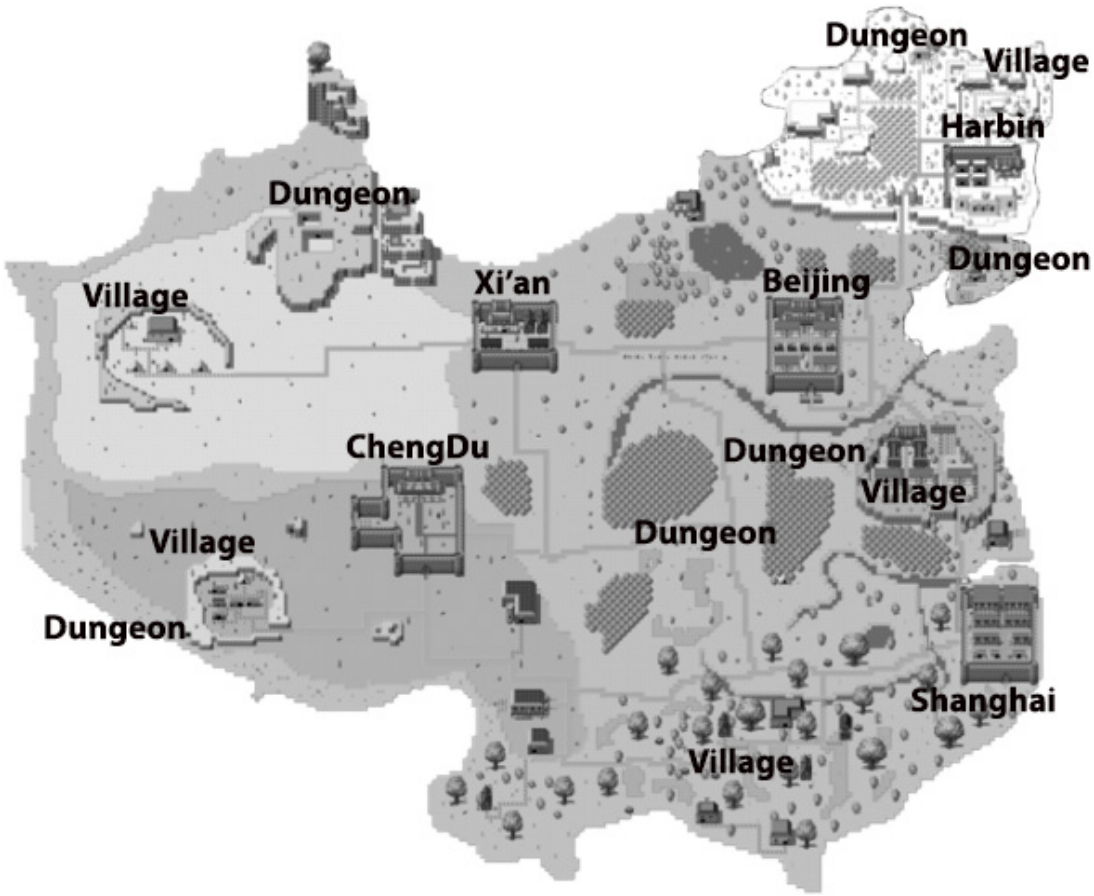
Game Design: Legend of the Dragon

The game used in the present study is called *Legend of the Dragon (龙之传说)* and was designed and built by the researcher using *RPG Maker MV*. In this game, students take on the role of an adventurer who sets out on a quest to aid the last dragon in China. Along their quest, students meet nonplayer characters (NPC) who provide information, present quests/tasks, and direct students towards the last dragon. The game world consists of five major Chinese cities (Beijing, Harbin, Shanghai, Chengdu, and Xi'an) and several fictional villages and dungeons placed in proximity to the cities. The game world was designed to resemble the shape of China with cities located in their approximate real-world locations (see Figure 2.1). To provide context for the size of the game, travelling from one side of the map to the opposite side of the map takes approximately 20 minutes and each of the cities, villages and dungeons lead the players into smaller sub-maps.

All players start in Beijing and after completing the initial tutorial are given the quest to retrieve a book in Xi'an. Once players retrieve the book, they learn that the last dragon in China is sick and they need to find three components (dragon blood, dragon

Figure 2.1

Over World Map of Legend of the Dragon



scales, and dragon bones) to concoct a potion that is believed to help the last dragon recover from the illness. Along the way, players are introduced to places and people that are culturally relevant to China. For example, the book that is sought in the first quest is located in an underground dungeon near Xi'an that houses Terracotta warriors.

Dialogue

To complete quests and in-game tasks, players have a set of cards, items, and skills at their disposal. These in-game features include a language support mechanism in

the form of a glossing system (W. Hong, 1997; Poole & Sung, 2016) that provides the Pinyin, a phonetic representation of Chinese characters. Glossing tools have been shown to support reading fluency (Shen & Tsai, 2010; Xie & Tao, 2009) and improve reading comprehension (W. Hong 1997; J. Wang, 2009, 2012; J. Wang & Upton, 2012) while also lowering reading anxiety (Zhao et al., 2013). It is important to note here that when designing the game, and more specifically when writing the dialogue, the researcher and a native-speaker of Chinese purposefully wrote the text so that in most instances potentially only one word would not be recognized. This is to say if a text was written that contained two potentially unknown words, the text was broken up into multiple text boxes (as shown in Figure 2.2) so that students were only given one potential new word in each dialogue box. This is important as past research shows that students should be able to read at least 95% of the vocabulary if they are expected to learn new vocabulary from the text (N. Liu & Nation, 1985). Word difficulty was determined from two sources. First, we created a master list of all the words used in the curricular materials developed by a publisher being used in the partner school (https://www.mandarinmatrix.com/). We focused on words that had been introduced in fourth grade or earlier for a majority of the words in the in-game dialogue, and used words introduced in fifth and sixth grade as our target vocabulary to be glossed. In some cases, words that did not appear in any of the grades, but were deemed important to the story were also included as a target vocabulary word to be glossed. The target vocabulary words were glossed to provide support in the form of pinyin. Finally, the students' sixth-grade teacher read through the texts to check for non-glossed words that students may struggle with.

For Chinese DLI students, many of the words in the game are review words from the in-class lessons and thus the students should be able to recognize a large portion of the target words orally. However, given the difficulty of recognizing Chinese characters, it is expected that many of the target words will not be recognized visually when first starting the game. Thus, when interacting with either a card, item, or skill in the game, players have the option to request the pinyin of the character (see Figure 2.2). The English translation is not given for a few reasons. First, and most importantly, English is not permitted in most DLI classrooms. Secondly, given the visual nature of the game, images and object representations should be sufficient to provide semantic knowledge for the characters.

Figure 2.2

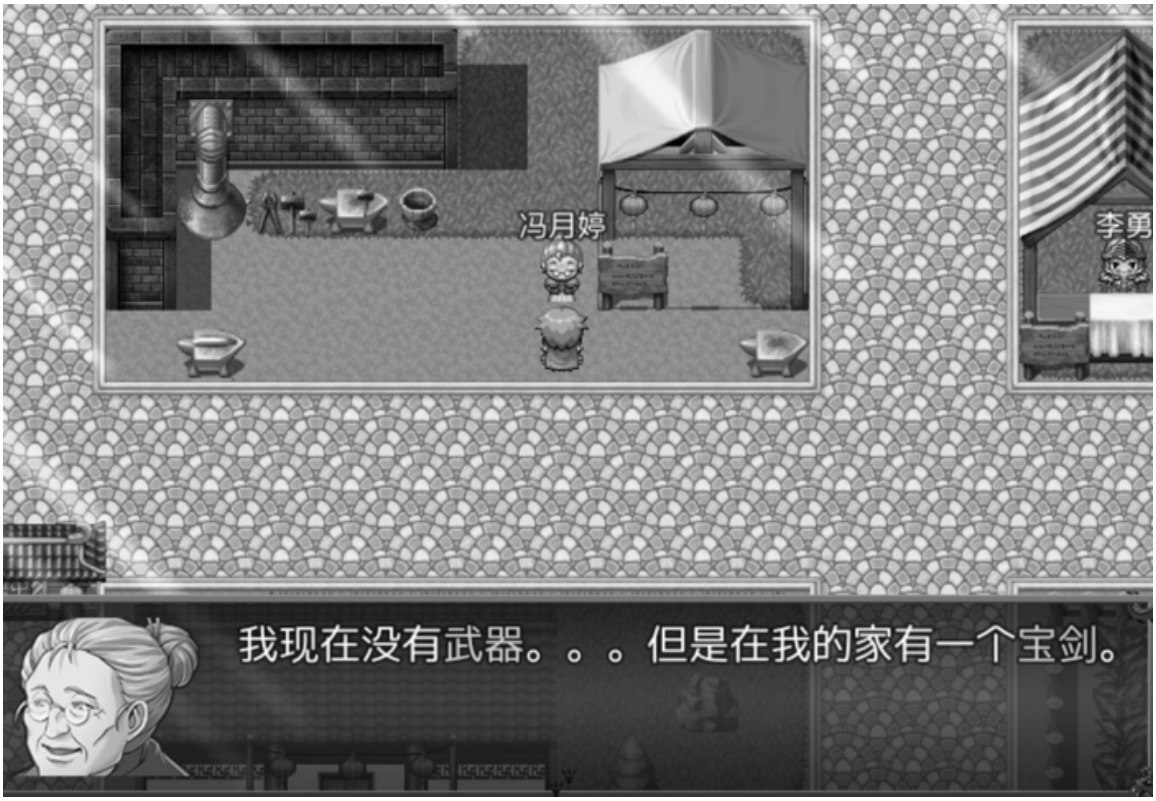
Glossing Tool

Quests

The quests involve a variety of puzzles, pick-up/delivery tasks, and enemy battling/taming activities. For example, one of the first quests that a player is given (see Figure 2.3) is to retrieve a sword for a weapon smith. The sword is located in the weapon smith's home, which is demarcated with a sign.

Figure 2.3

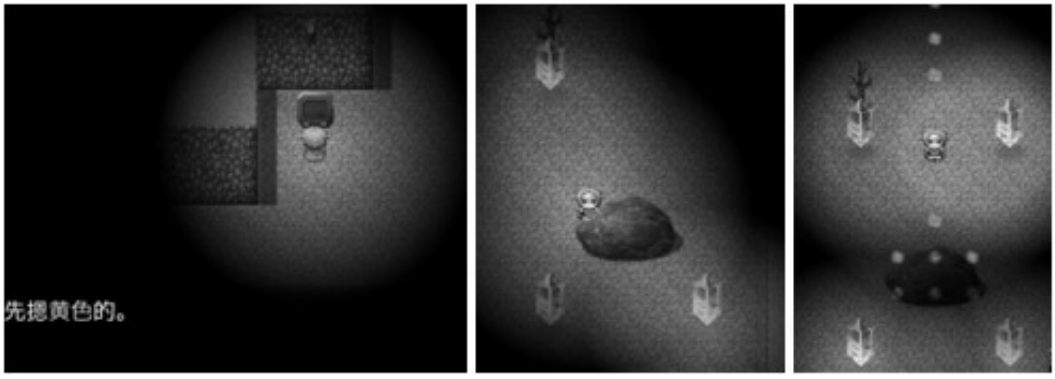
Sword Retrieval Quest



Players must also solve puzzles that utilize their language skills. For example, in the puzzle below (see Figure 2.4) students must read the text from a sign that indicates a sequence of colors and then activate the orbs in the correct order to open a gate.

Figure 4

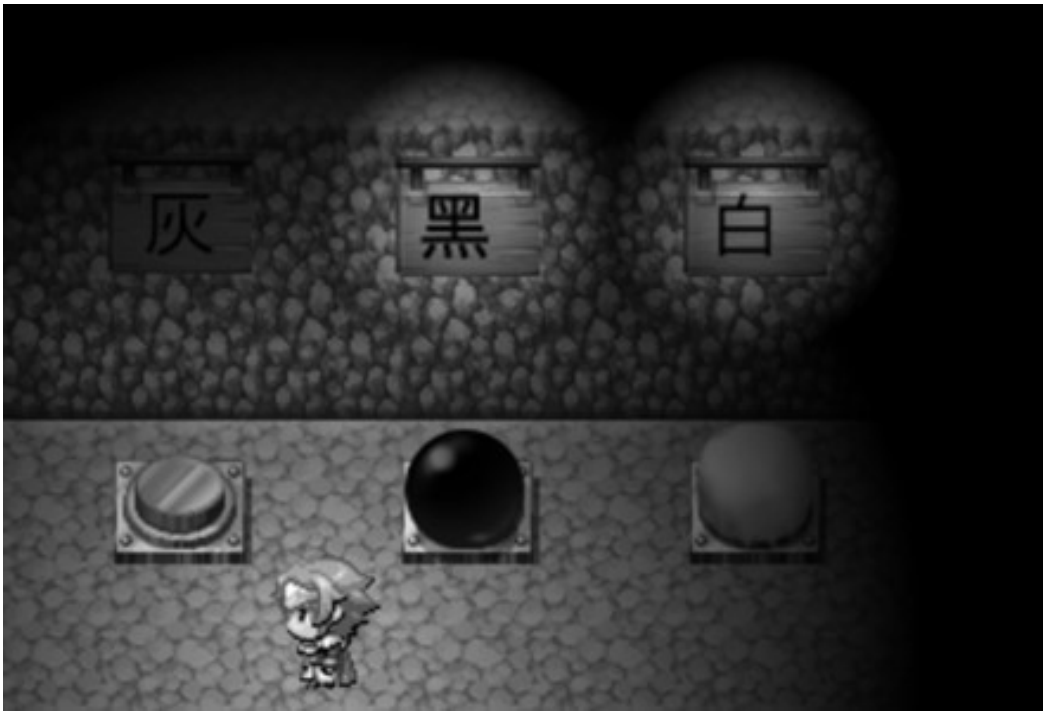
Sequencing Quest



In a similar puzzle (see Figure 2.5) the player must match the color of the boulder with the correct sign to open a gate to the final boss.

Figure 2.5

Matching Quest



The player completes the quests by interacting with a series of text-based dialogue that direct the players in the correct direction and completing mini-puzzles or language-based tasks. These quests were designed to make the text meaningful, and thus promote a learner's attention to what is being read (J. Lee & VanPatten, 2003). Finally, when a player is given a quest, a quest bar at the upper right-hand corner of the screen appears (see Figure 2.6) with a summary of the quest and important names and items highlighted.

Figure 2.6

Quest Bar

This quest bar was seen as essential, as past experiences with the elementary classes involved in this study (Poole et al., 2018) indicated that young learners can become easily distracted and forget the task at hand.

Battle System

While exploring the game, players must navigate through several baddies spread throughout the fictional world. Baddies are one of 20 different animals and each baddie has a special type of attack, strength, weakness and preferred foods. This information can be found on baddie cards that players collect by either defeating or taming a baddie. Players can also purchase a subset of baddie cards at a card shop in each of the cities. These baddies range in difficulty from 1 to 20 with 20 being the most difficult. When players come into contact with a baddie they have two options: battle or tame. If they choose to tame a baddie they can do so by giving the baddie it's preferred food. To battle a baddie, the player again has two options: attack or summon a helper. All players have a base attack ability, but the base attack ability is very weak. Even when collecting swords, armor, and/or rings the players base attack is barely enough to defeat lower-level baddies. This was a conscious game design choice made to encourage use of baddie cards that allow players to summon an animal to fight by their side. It is important to note that players must collect at least five baddie cards before they can learn the skills to summon a particular animal. Once the skill is learned, they need a card each time they summon an animal. This was done to encourage the use of different animals. Rather than simply summon one animal every time, players need to summon different animals as their cards become available. Further, certain animals' strengths are better when fighting baddies with similar weakness. This game mechanic was designed to encourage the students to read the information on the animals' cards in order to learn which animals should be summoned in certain battles. Thus, regardless of whether students engaged in battle or if

they decided they wanted to tame a baddie, they would need information on the cards.

Supplemental Materials Design

Supplemental materials in the form of a paper-based workbook written in Chinese were specifically designed to support the learning in the game. The workbook was designed to promote character recognition, oral discussions around game items and places, and note-taking in terms of in-game exploits. The workbook provided two to three language activities for students to complete in addition to playing the game. These language activities included a character writing task, which asked students to write commonly seen characters in the game three times. Past research has noted the positive effects of writing practice on character recognition skills (Guan et al., 2011; Tan et al., 2005). A second task asked learners to match images to characters to further promote character recognition skills. Finally, discussion questions asked learners to use the targeted vocabulary words to discuss features of the game (e.g., Where do you find the battle cards?), game items (e.g., What sword color does the most damage?), gameplay strategies (e.g., Which partner animal is best?), and quests (e.g., What quest does Sima qian give?). The workbook also contained a printed world map that students were encouraged to keep notes on. The workbooks were collected at the end of the project and were reviewed based on completion of exercises. Approximately 82% of the students completed the exercises in the workbook each day.

Teachers' Role and Classroom Implementation

During the integration of the game into the classroom, the researcher took on the

role as the primary instructor, and the teacher of the class took on a support role. The teacher introduced the project to their students, and then the researcher directed students into groups, provided support while students played the game, and provided an overview of workbook activities each day. During gameplay sessions, students were divided into four groups of five students each. Two groups first started working on the activities in the workbook while the other two groups played the game. After approximately 25 minutes the groups switched, so all students received 25 minutes of gameplay and 25 minutes of workbook time. During the first week of the intervention, the researcher provided an overview of the tasks in the workbook. In the following weeks, students were able to complete a majority of the assigned workbook tasks without further support. In terms of gameplay, in the first week, while students were in the tutorial area, the teacher and researcher both walked around the room, from group to group to provide technical and basic gameplay support while students played the game. In the following weeks, the researcher and teacher interacted with students while they played by answering student questions about the game, asking questions about player progress, providing hints, and reminding learners of quest tasks as they played.

Research Design

This study took place in a sixth-grade Chinese DLI classroom located in a rural town in the western U.S. All but four of the students in this class started the Chinese DLI program in first grade. Of the four who did not, two were native speakers and the other two had experience with Chinese in previous schools. The teacher was a second-year

Chinese DLI teacher who had expressed interest in using games in her classroom. I personally have worked with this particular class in two past projects and thus was familiar with the students, and likewise they were familiar with me and gameplay in the classroom. While following this class for the last 5 years, I have noted that they have strong oral skills, but tend to struggle with recognizing Chinese characters and subsequently reading comprehension. This is similar to past studies that have noted that Chinese DLI students' reading comprehension skills tend to lag behind their peers studying other languages in similar immersion programs (Burkhauser et al., 2016; Watzinger-Tharp et al., 2018).

Participants and Setting

Data were collected in two sixth-grade elementary Chinese DLI classroom settings. The two classrooms had 19 and 21 students respectively. Stratified randomization with gender as a factor was used to assign learners to each classroom at the beginning of the school year. Students ranged in age between 10 and 12 years old (mean = 11.05). Four students (1 in Class A, 3 in Class B) did not sign the consent form (see Appendix C). In addition, two students in the Class B did not take the pretest, and one student in both classes did not take the post test. Thus, although I had consent forms for 36 students, there was only complete data from 32 students, see Table 2.1.

Sixth-grade learners were chosen for a few reasons. First, for L2 learners to read and learn from an L2 text, it is suggested that they know about 95% of the vocabulary in the text (Liu & Nation, 1985). After conducting past studies with the second, third, and fourth grade classrooms, working extensively with the current sixth-grade classrooms on

Table 2.1

Students Included in Data Analysis

Class	$N_{students}$	Survey data
A	20 (10F, 10M)	20 Pre, 19 Post
B	16 (9F, 7M)	14 Pre, 15 Post
Total	36 (18F, 16M)	34 Pre, 34 Post

Note. Only 32 complete cases.

other projects, and talking with the teachers of each grade, a mutual decision was made (between the teachers and the researcher) that the text difficulty was most suitable for the sixth-grade classes. Second, the game utilizes pinyin as a means of support for unknown words. In this particular DLI program, pinyin is not introduced until the third grade. Third, it was important to include classroom vocabulary within this game. In the first three grades of the DLI program, Math is the major subject taught in Chinese thus limiting the vocabulary and subsequently the types of quests that could be used in the game. The sixth-grade students learn science in Chinese and cover topics that include: landforms, heredity, magnets, electricity and matter among other topics. Finally, in past studies with younger learners, there was a large learning curve when introducing new computer programs. The sixth-grade classrooms in this study had steady computer use since the third grade and thus computer proficiency was not major a concern.

Recruitment

I have worked with this Chinese DLI program for 5 years. In that time, I have developed relationships with most of the teachers and the administration. The sixth-grade teacher expressed an interest in using the game in their classroom.

In terms of recruiting students, I presented the project in the classroom two weeks before administering the pretests and explained that although playing the game is a part of the class curriculum, students do not have to have to participate in the research aspects of the project (affect surveysetc.). Participants were then given a consent form to take home to their parents and given one week to sign the form and return it to the classroom. Participants whose consent forms were lacking either parental or student signature were allowed to play the game, but their data collected within the game were removed after gameplay. Furthermore, these students were not given an affect survey. These students were still given a pre- and post-vocabulary and reading comprehension test per request of the teacher, but these data were not used for research purposes.

Procedures

On the Friday prior to gameplay, participants completed the paper-based pre-assessment. In the sixth-grade classroom, there is a 1:1 computer to student ratio. However, *RPG Maker MV* games are not compatible with Chromebooks. Thus, the researcher brought in eleven MacBook laptops to play the game. In order to maintain a 1:1 ratio, the class was divided into four groups of five students each, with a mixture of high and low proficiency learners as identified by the instructor of the class. Two of the groups played the game while the other two groups completed the workbook. After 25 minutes, these groups switched.

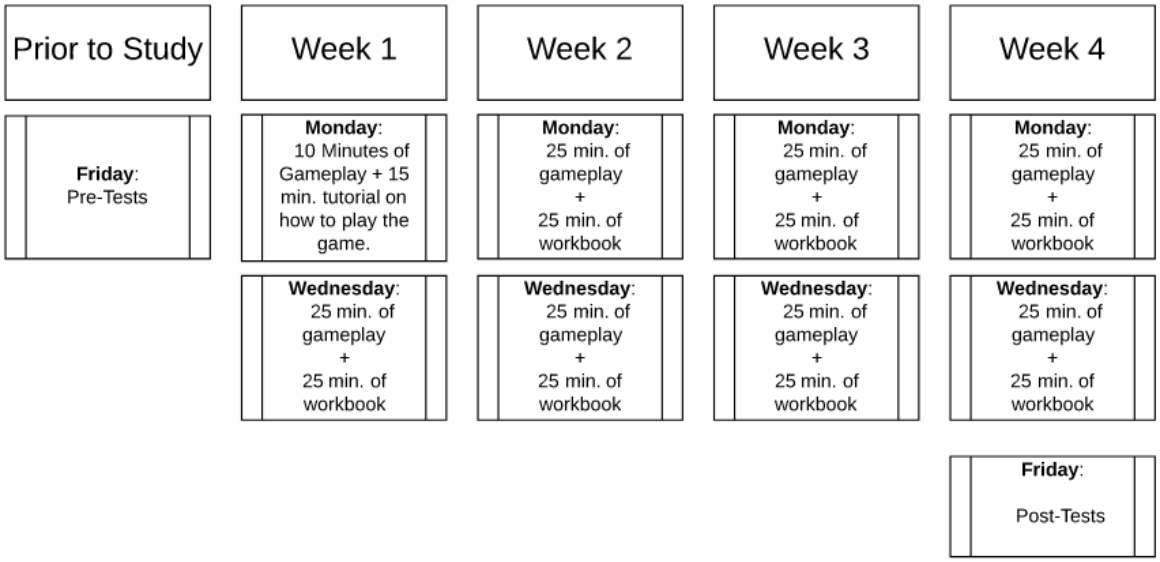
On the first day of gameplay, the participants were given a brief tutorial on how to play the game via a whole-class demonstration. The classroom teacher then illustrated how the game related to their current studies by telling the students that vocabulary in the

game was primarily comprised of review words and that the game was similar to their current in-class readings in that they were reading about students who were traveling to China. Similarly, in the game the students would be going to China and could explore the same places that the characters in their books were visiting (e.g., Beijing, Xi'an, Chengdu).

Participants played the game for 50 minutes (two 25-minute sessions) and completed supplemental materials for 50 minutes (two 25-minute sessions) per week over the course of a four-week period (once on Monday and once on Wednesday). Participants played the game for approximately 3 hours over the four weeks. Students were administered the paper-based post-assessment two days following the final game play session. Figure 2.7 illustrates the timeline for the study.

Figure 2.7

Research Schedule



Data Collection and Analysis

Data for the present study came from five sources: pre- and post-vocabulary assessments, pre- and post-reading comprehension assessment, and audio recordings during gameplay.

Quantitative Data

Vocabulary Assessment

The pre- and post-vocabulary assessment consisted of 45 words that could be found in the game. All of the vocabulary words selected were words that could be checked via the pinyin glossing system. It should be noted that this is not a comprehensive list of the vocabulary words that could be checked. There were 326 words that learners could potentially be exposed to via the glossing system while playing the game. Assessing all vocabulary within the game would not be feasible given the time constraints associated with the DLI program. Thus, target vocabulary was selected after discussions with both the teacher and the Chinese native-speaker who helped write the dialogue. Words that were both deemed important to the overall storyline in the game and relatively unknown to the students were added to the assessments. On the vocabulary assessments, learners were presented with the character and then were asked to declare if (a) they know the word, (b) they think they know the word, or (c) they are just guessing. Next, learners were prompted to enter the pinyin and the English translation. The vocabulary assessment can be seen in Appendix A. These assessments were scored by awarding 1 point for correct pinyin, and 1 point for correct English definitions. Half

points were awarded for partial answers. For example, if a learner correctly identified the correct pinyin or English for one of the characters but not both, half points were awarded. Awarding partial points for vocabulary knowledge was viewed as valuable given past research that has noted the non-linear trajectory, and partial accumulation of vocabulary knowledge that occurs through incidental learning while reading L2 texts (Grabe & Stroller, 2002; Horst, 2005; Pigada & Schmitt, 2006).

Reading Comprehension Assessment

This was adapted from the Youth Chinese Test (YCT) (http://english.hanban.org/node_8001.htm), which is an official Chinese proficiency assessment developed by the Confucius Institute and used regularly by the Chinese DLI program that served as the context for the present study. The reading comprehension assessment consisted of 10 items. See Appendix A for the items. Although the format of the reading comprehension assessment was adapted to reflect the YCT, the content was adapted to reflect text that the learners might see in the game. To reduce the priming effect on the reading comprehension assessments, items were randomized in both the pre- and post-assessments. Further, although the sentence structures remained the same, key vocabulary words, and thus the answers, changed from pre- to post-assessments. For example, one of the questions first states that bats like to eat fruit and then asks what would a bat like to eat with three options. On the post test, the question states that wolves like to eat vegetables and then asks what would a wolf like to eat again with three options.

Previous Gaming Experience

On the pre-assessment, students were asked two questions about their previous gaming experience. The first question asks if students play video games at home. The second question asks how often students play at home with possible answers ranging from “I have only played a few times in my life” to “Everyday” on a 5-point Likert scale.

Qualitative Data

While students played the game, audio recorders were used to capture in-class dialogue around the game. Informed consent was not obtained for the teacher of this class, and so only the support provided by the researcher was analyzed. While the teacher of the class did engage with students and provide support, it is important to note that the type of support provided by the teacher and researcher, myself, are different given my familiarity with the game. I am an avid gamer and I have both taught and conducted research with games in multiple occasions in other contexts. Further, given that I designed and built this game, I am very familiar with every quest, map, and trick to the game. It should also be stated that I view games not as stand-alone learning environments, but rather as environments that can be leveraged as both highly contextualized learning moments and for meaningful interactions between students and teachers. Thus, while I circulated the room looking for students who may need help, I was also looking for opportunities to discuss in-game exploits and to encourage students expand on statements that they made around the game. Through conversations with the teacher throughout this project I shared my beliefs and perspectives about how games should be used and we engaged in many discussions about when and how to provide

support. Thus, the teacher did adapt some of my strategies for interacting with and providing support to students as they played the game.

Finally, it should be noted that some audio was incomprehensible. When this occurred, a second audio recorder was checked for clarity. However, there were some researcher-student interactions that were not captured. Table 2.2 provides an overview of all data collected in this study and how the data were used.

Table 2.2

Data Sources

Data source	Time	Description
Knowledge/skills		
Vocabulary assessment	Pre- and Post-	This measures student knowledge on 45 vocabulary words that are found in the game.
Reading comprehension assessment	Pre- and Post-	There are 10 reading comprehension questions that use vocabulary and sentence structures from the game.
Background		
Previous gaming experience	Pre-only	This consists of two questions, one that asks if the participants play video games and second that ask how often.
Audio Files		
Audio recordings of teacher talk	During intervention	Teacher interactions with students were recorded during gameplay.

Data Analysis

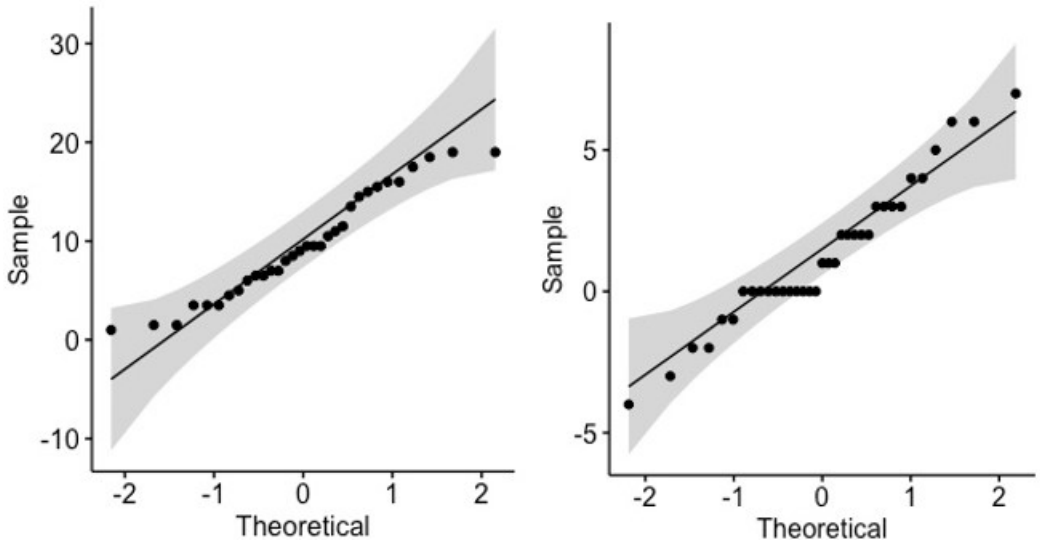
Quantitative Data

The first research question investigates whether learning occurred during the intervention. To answer this question, paired samples *t* tests were used to compare pre- and post-vocabulary and reading comprehension scores. Before running *t* tests, it is

important to check for significant outliers in the difference between pre- and posttests. In addition, it is important to check that the difference of pairs follows a normal distribution. There were no outliers in the difference scores for the reading comprehension assessments or the vocabulary assessments. The distributions of the difference scores for both assessments also satisfied the normality assumption as assessed by the Shapiro-Wilk test (Vocabulary Assessment, p =.149, Reading Comprehension Assessment, p = .223) and by examining the QQ plots for the difference scores (see Figure 2.8). We used Cohen's d, a standardized measure of the differences between the means, to calculate the effect size. Effect size provides a description of magnitude of the observed effect that is independent of sample size (Fritz et al. 2012).

Figure 2.8

QQ Plot for Vocabulary Assessment on Left, QQ Plot for Reading Comprehension Assessment on Right



The second research question investigates whether or not the learning gains are

dependent on students' prior gaming experience. A step-wise linear regression analysis in R (R Core Team, 2017) was run to determine if prior gaming experience affected learning gains. To ensure there were no violations of assumptions of normality, linearity, multicollinearity, and homoscedasticity, preliminary analysis was first conducted and no violations were identified. For both models, vocabulary and reading scores were used as the dependent variables. Then, pre-vocabulary and reading comprehension scores were entered into the model followed by the previous gaming experience variable. An ANOVA was then used to determine if adding prior gaming experience contributed significantly to either of the outcome scores.

Qualitative Data

The third research question explores the type of support provided by instructors in the present study. The researcher of the present study engaged in multiple rounds of coding of the utterances. First, the researcher did a round of open-coding, where they looked for patterns, themes, and categories in the data (Saldaña, 2015; Strauss & Corbin, 1998). In the second cycle of coding, the researcher employed axial coding (Patton, 2014; Strauss & Corbin, 1998), focused on organizing the codes into categories that best explained the types of supports. This round of coding involved iterative cycles of review and revision to achieve saturation of categories. At the end of this second cycle, ten codes were organized into seven categories.

To calculate inter-coder reliability, an additional person, a Native-Chinese speaker, was trained on the coding procedure and provided with the codes and definitions (see Table 2.3).

Table 2.3

Description and Example of Codes

Code	Description	Example
Quest management	Direct learners towards to the next quest or next part of a quest.	但是你先得把地图还给那个人。 [[But you need to take the map back to this person first.]]
Battle strategy	Discuss strategies related to winning a battle.	你不应该用猴子，应该用蝙蝠。[[You shouldn't' use the monkey, you should use the bat.]]
Encouragement	Encourages a player to explore the game or read a text on their own.	这个书很重要，你必须看懂了，才知道下一步是什么 [[This book is really important, you have to understand to be able to move on to the next part.]]
Technology	Shows player how to play the game or helps with a technical problem related to the game.	你得保存. [[You must save the game.]]
Confirmation	Confirms a question or belief about the game.	猪，对，你要抓三只猪。 [[Pig, Right, You need to catch three.]]
Talk around the game	Discuss current status of the game, object in the game, and/or results of an event in the game.	他马上要赢了。他很快就赢了。 [[He's going to win, soon, he'll win!]]
Linguistic	Helps student read a text or provides a translation to a word.	战斗，战斗就是打仗。 [[Battle, battle is just like fighting.]]

After receiving training, the coder was provided with 20% of the utterances to code. Cohen's κ =.74, indicating a "moderate" level of agreement (McHugh, 2012). Any discrepancies between the codes were then discussed between the two coders and applied to the remaining utterances.

Results

Descriptive Statistics

In terms of learning, the vocabulary assessment had a range of 0 to 90, and the reading comprehension assessment had a range of 0 to 10. On average, students reported

a score of 18.31 on the pre-vocabulary assessment, which is approximately 9 out of 45 words (a total of 2 points were given for each word). On the post-vocabulary assessment students scored an average of 27.95 points or 14 out of 45 words. Neither the pre-vocabulary nor the post-vocabulary assessment violated assumptions of normality. Although it should be noted that the pre-vocabulary assessment was more skewed and kurtotic than the post-assessment. As for the reading comprehension assessment, on average students scored 3.41 on the pre and 4.72 on the post. Neither of these assessments violated the assumptions of normality, see Table 2.4.

Table 2.4

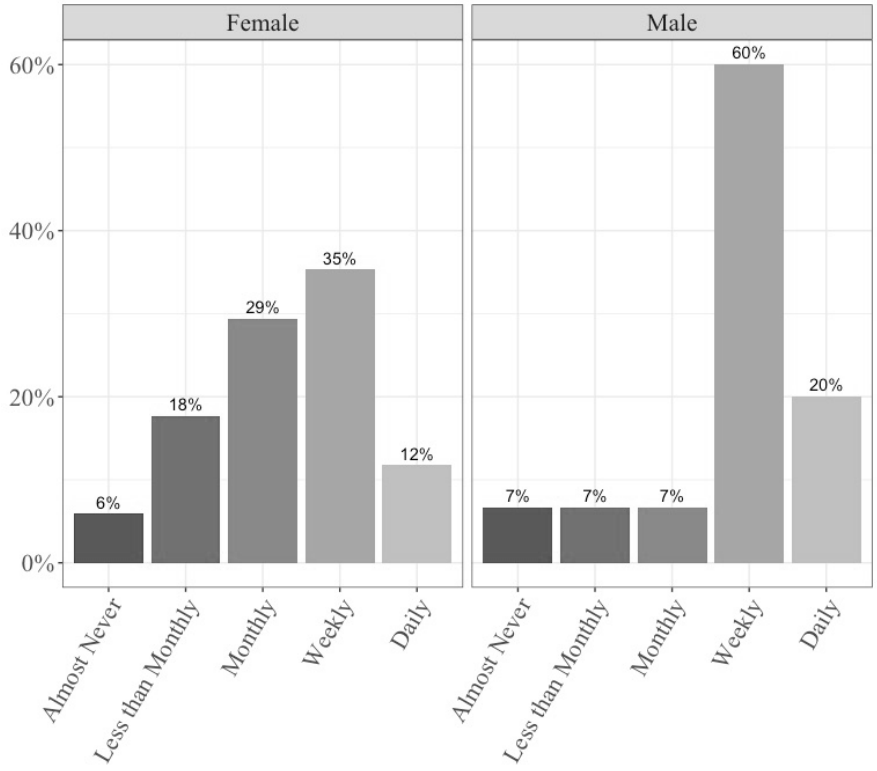
Vocabulary and Reading Comprehension Descriptives

	Pre					Post				
Measures	*M*	*SD*	α	Skewness	Kurtosis	*M*	*SD*	α	Skewness	Kurtosis
Vocabulary (*n* = 32)	18.31	14.38	.95	1.45	2.20	27.95	17.97	.96	1.14	1.08
Reading comprehension (*n* = 32)	3.41	2.56	.74	0.39	-0.98	4.72	2.43	.60	0.15	-0.78

On the pre-survey, students were asked to report their current gaming practices. This was seen as important as it may predict who learns from the game. These survey findings are presented in Figure 2.9 illustrating differences between boys and girls. About 80% of the boys play video games at least once a week or more, while only 47% of girls in the class reported playing the same amount. Although this is a stark difference, it should be noted that all students reported playing games.

Figure 2.9

Game Experience by Gender



Research Questions

Research question 1 asked, *"Do students who play the digital game show learning gains in vocabulary or reading comprehension, as measured by the pre-and-post assessment?"* This research question is concerned with the effect of the game on student learning. To answer this question, the pre- and post-assessment scores for both the vocabulary and reading comprehension measures were compared using *t* tests, see Table 2.5. Students reported significantly higher vocabulary scores on the post-assessment (M = 27.95, SD = 17.97) than on the pre-assessment (M = 18.31, SD = 14.38), which indicates that learning did occur as a result of the intervention, $t(31) = 9.99$, $p < .001$). The effect

Table 2.5

Learning Measures

	Pre				Post					
Measures	*M*	Med	*SD*	*n*	*M*	Med	*SD*	*n*	*t* test	Cohen's D
Vocabulary	18.31	16.00	14.38	32	27.95	23.50	17.97	32	9.99***	0.45
Reading comprehension	3.41	3.00	2.56	32	4.72	4.50	2.43	32	3.22**	0.54

Note. Vocabulary Scale from 0-90. Reading scale 1-10. Both effect sizes are medium.

** $p < .01$.

*** $p < .001$.

size was .45, a medium effect size for educational interventions such as this one (Cohen, 1988). In other words, after seven 25-minute gaming sessions along with supplemental workbooks, the students in this sample gained, on average, about 10 points of vocabulary knowledge. This translates roughly into a five-word increase.

Similarly, students reported significantly higher reading comprehension scores on the post-assessment ($M = 4.72$, $SD = 2.43$) compared to the pre-assessment ($M = 3.41$, $SD = 2.56$), indicating that students' reading comprehension also increased significantly, $t(31)=3.22$, $p < .01$). The effect size was .54, a medium effect size (Cohen, 1988). Students in this sample gained, on average, slightly more than 1 point on the reading comprehension assessment after the intervention.

Research question 2 asked, *"Do students who have prior gaming experience see greater gains in vocabulary or reading comprehension?"* To further explore whether or not prior gaming experience was associated with learning gains, a step-wise linear regression was run with both reading and vocabulary gains as dependent variables. For both of these models, pre-assessment scores were first entered into the model and then

the prior experience variable was entered in a following model. As Table 2.6 illustrates, adding prior experience to a model with pre-vocabulary scores did not significantly account for more variance in reading gains as assessed by ANOVA (p = .583). Prior gaming experience only added a .007 increase in the total amount of variance in reading gains accounted for by the model.

Table 2.6

Linear Regression of Prior Gaming Experience on Reading Gains

Independent variables	Reading gains (Model 1)	Reading gains (Model 2)
Pre-reading score	-0.506***	-0.499**
	(-0.772, - 0.240)	(-0.769, -0.229)
Prior gaming experience		0.178
		(-0.449, 0.804)
Constant	3.036***	2.385
	(1.907, 4.164)	(-0.178, 4.948)
Model fit (ChiSq)		.583
Observations	32	32
R^2	0.316	0.323
Adjusted R^2	0.293	0.277
Residual std. error	1.939 (df = 30)	1.961 (df = 29)
F statistic	13.875*** (df = 1; 30)	6.932** (df = 2; 29)

** p <.01
*** p < 0.001

Similarly, Table 2.7 shows that prior gaming experience was also not a significant predictor of vocabulary gains when controlling for pre-vocabulary scores. Adding prior gaming experience only increased the total variance accounted for by the model by .001 and was not significant as assessed by ANOVA (p = .817). The implications for these findings will be further explored in the discussion section.

Table 2.7

Linear Regression of Prior Gaming Experience on Vocabulary Gains

Independent Variables	Vocabulary Gains (Model 1)	Vocabulary Gains (Model 2)
Pre-reading score	0.209**	0.210**
	(0.096, 0.322)	(0.094, 0.325)
Prior gaming experience		-0.125
		(-1.626, 1.376)
Constant	5.811***	6.241*
	(3.188, 8.435)	(0.433, 12.049)
Model fit (ChiSq)		.871
Observations	32	32
R^2	0.303	0.304
Adjusted R^2	0.280	0.256
Residual std. error	4.631 (*df* = 30)	4.708 (*df* = 29)
F statistic	13.070** (*df* = 1; 30)	6.336** (*df* = 2; 29)

* $p < .05$.
** $p < .01$.
*** $p < 0.001$.

Research question 3 asked, *"What supports, if any, do students need from instructors when integrating the digital game into the classroom?"* The third research question concerns the types of support that students may need as a result of playing a digital game for L2 learning purposes in a classroom setting. This question was answered by analyzing the support given during gameplay sessions by the researcher of this study. After coding the data, seven types of support were identified during the four weeks of gameplay. Table 2.8 provides an overview of the frequency and percentages of the categories identified in this study. A description of the supports and examples are provided below.

Table 2.8

Types of Support Provided during Gameplay

Supports	Frequency	%
Quest Management	98	28.1
Battle Strategy	71	20.3
Talk Around Game	70	20.1
Encouragement	53	15.2
Technology Support	53	15.2
Confirmations	23	6.6
Linguistic	13	3.7
Total Utterances	348	

Note. Some utterances contained multiple codes and thus the total utterances do not reflect the total number of codes.

Quest Management

The most frequent type of support provided was in the form of *Quest Management*. This support either prompted students to start a quest, directed them to the next part of a quest, or reminded them of their progress in a quest. This form of support was either prompted by the student asking what they should do next or by the researcher when he noticed that a student was struggling. In the example below, a student was grabbing items from around the starting city and picked up a flower. This student wanted to know what the flower did.

Student: 它会做什么？ [[What does this do?]]

Researcher: 你问这个人在下面，你问这个人她会告诉你它会做什么. [[Ask this person down here, you ask this person and she will tell you what it does.]]

The researcher took this opportunity to direct the student towards the start of a quest in the tutorial. The flower that was being collected by the student was an item that was

required to complete a quest. In a later example, on day seven, a student was looking for an important person in one of the quests and was unable to find them. The student asked the researcher for help finding the person.

Student: 所以我不可以找到这个人 [[So, I can't find this person]]

Researcher: 他在这个地方 [[He's in this place.]]

Student: 这儿？ [[Here?]]

Researcher: 对。你得绕过来. [[Right, so you need to go around.]]

In this example, the researcher pointed to a location on the student's workbook map, and then drew a path with his finger to illustrate how the student would need to leave the city and go around the walls to reach the dungeon where the person of interest was located.

Battle Strategy

The second most frequent type of support came in the form of *Battle Strategies*. These were instances when the students either wanted to discuss strategies for winning a battle, or if the student was about to lose a battle and the researcher provided a hint or insight. For example, when walking by a student, the researcher noticed that the student was about to lose a battle to a spider that was a much higher level than the student's character.

Researcher: 你死了？ [[Did you die?]]

Student: Yeah.

Researcher: 他喜欢吃虫子。你有虫子吗？ [[He likes to eat bugs. Do you have bugs?]]

Student: 没有 [[No, I don't have any.]]

> Researcher：他很厉害，你不应该跟他打 [[He's strong. You shouldn't fight with that one yet.]]

Here the researcher provides a hint that spiders like to eat bugs, so that next time when the student encounters the spider, she can feed it bugs rather than engage in battle. The researcher reminds the student that she should fight with lower-level baddies. In another example, the researcher suggests that a student might want to return to Beijing to pick up some equipment before they engage in battles.

> Researcher: 有可能你要回去北京。买一点东西，买一个宝剑什么的 [[You might want to go back to Beijing to buy some stuff, buy a sword or something.]]

Talk Around the Game

This category included moments when the researcher was asking about either current progress or activities that a student was engaging in while playing the game or comments about the status of the game. Comments about the game include talking about items that the learners acquired, character levels, or the results of battles (e.g. 你真厉害 [[You're awesome!]]. For example, on the last day of the study, one student figured out that purple items were the best items in the game, but this particular student was unable to find the purple sword. While searching in Chengdu, this student loudly notifies the researcher that the purple sword is not there.

> Student: 他们没有 [[They don't have it here.]]
>
> Researcher：这个就关门了， 你去西安。 西安应该有。[[That one is closed. Go to Xi'an. Xi'an should have it.]]
>
> Student：但是他们没有紫色的 [[But they don't have the purple one.]]
>
> Researcher：是那儿，应该有。对不起，哈尔滨有， 我忘了。 [[It is there, they should have it. Oh sorry, it's in Harbin, I forgot.]]

As the student continues to search for the sword, the researcher starts making suggestions on where the sword might be and first suggests Xi'an. The student remembered that it was not there, this prompts the researcher to suggest Harbin.

Encouragement

The *Encouragement* category occurred when the researcher simply suggested that the students carefully read a text or explore a part of the game. This happened when the researcher felt that a particular text was really important. For example, one student was confused about what she should do next and asked the researcher for help. Upon further investigation, the researcher discovered that the student had acquired the Dragon book, which sets up much of the story, but had not opened it. So, the researcher helped the student open the book and then encouraged her to carefully read what was inside.

> Researcher: 所以你现在，你看这，你有龙之书，你要看他，好好看。 [[So now, you have to read this, you have the Dragon Book, you have to read it. Carefully read it.]]

In another example, a student was trying to complete a quest that required the student's in-game character to locate a small hamster in a dungeon. The student wanted to know where the hamster was, and the researcher simply told the student to look around.

> Researcher: 你找一找吧，你进去看一看。[[You have to search for it. You should go in (the dungeon) and look around.]]

Technology

The technology code was used to categorize moments when either the student had a problem with the user interface or a student needed help with the game controls. For example, in the first few game sessions there were questions about saving the game, how

to use the cards, access the control menu, how to put equipment on, and how to look up vocabulary with the glossing system. This led to the researcher showing students how to restart the game, point out where things were in the menu, or simply telling the student to use the "control" key to look up an unknown word.

Confirmations

This category includes moments when students wanted confirmation about what they were doing. For example, one of the quests in the game has students rearrange a set of tombstones to collect a prize. In the dialogue the word tombstone is glossed, thus provides pinyin for the student. The student did not know this word before playing, but given the context of the game, wanted to confirm that the three tombstones in the game were in fact tombstones. The student first asked his classmates, and then confirmed with the researcher.

Student: 这个是墓碑？ [[This is a tombstone?]]

Researcher: 是的，这三个都是墓碑 [[Yes, those three are all tombstones.]]

Linguistic Support

The least used type of support is *Linguistic* support. This came when students had direct questions about unknown words. These were usually words that were a part of the game rather than part of the students' in-class vocabulary. For example, one student asked what 战斗 [[*Battle*]] meant, to which the researcher said, 战斗，战斗就是打仗。[[Battle, battle is the same as fight]]. Similarly, another student wanted to know what 蝎子 [[scorpion]] was, to which the researcher said, 蝎子是，看，它的尾巴。[[Scorpion

is, here look, it's tail is …]] while mimicking a scorpion's tail with his hands. The implications for these types of supports will be discussed in the following section.

Discussion

While the L2 field has seen an increase in studies on games and L2 learning, there are very few studies that focus on what it means to design and integrate games into L2 classroom instruction. There are even fewer studies that look at L2 games that teach Chinese in elementary classrooms. In the present study, the researcher first designed a game to promote Chinese reading comprehension and vocabulary learning in sixth grade DLI classrooms and then implemented it in a classroom.

The first research question explored whether or not students learned by interacting with the game and supplemental activities in the classroom. The results indicated that students had significant gains in vocabulary after participating in the project. On average, students gained approximately 10 points on the vocabulary knowledge assessment which roughly translates to five words. While five new words in this study may not seem like a large gain over the course of four weeks, these were similar to other studies exploring vocabulary gains in an educational game. For example, Chen and Yang (2013) found a significant increase in vocabulary knowledge after 1.5 hours of gameplay and notetaking. The significant difference was equivalent to a gain of two new words. In another example, Alyaz and Genc (2016) allowed preservice teachers to play an educational game for the last part of class for 8 weeks. They also reported significant vocabulary gains equivalent to 5 learned words over the 8-week period. It should be noted that both

of these studies were working with adult learners of English as a second language.

In addition, students in the present study were only assessed on 45 vocabulary words. Given that students were exposed to approximately 500 texts on average over the course of 4 weeks (see Chapter 3), it is likely that they were exposed to many more words, and may have learned words that were not among the 45 assessed. Additionally, upon further analysis in Chapter 3, it was determined that five of the words assessed were only seen by a small portion of the students in the game, thus making it unsurprising that they were not learned.

Finally, students in the present study were not specifically told to focus on vocabulary learning as was done in other studies investigating Chinese vocabulary learning in a virtual environment (e.g., Hsiao et al., 2017; Lan et al., 2015). Thus, the increases in vocabulary knowledge can be seen as being learned incidentally, or as "a by-product, not the target, of the main cognitive activity" (Huckin & Coady, 1999, p. 182). Incidental vocabulary learning that occurs in reading has been argued to be a better approach for increasing vocabulary knowledge because the words are contextualized in text and learners have the opportunity to improve reading skills while learning vocabulary (Huckin & Coady, 1999; Krashen, 2004). However, tracking incidental vocabulary learning while students read a text can be difficult given that learners often report gains in partial knowledge of a word (Horst, 2005; Pigada & Schmitt, 2006), much like in the present study. While research has shown that vocabulary can be increased incidentally through reading, it must be done through extensive reading (Krashen, 2004). Contextualizing the vocabulary gains in terms of past research, student exposure to in-

game text, and the context of the study highlight the challenges of designing an educational game. On one side, it was seen as important to allow the learner autonomy and freedom in how they play the game. Autonomy in gameplay has been linked to enjoyment (K. Kim et al., 2015), further by giving the learner autonomy they are effectively allowed to explore and experiment with the game, which can lead to meaningful opportunities to learn. However, by giving the learner autonomy via an open-world design, it becomes difficult to control what text and vocabulary that learners are exposed to.

In addition to increases in vocabulary, the present study also found significant increases in reading comprehension scores. On average, students scored approximately 1.3 points higher on the post-assessment than on the pre-assessment and a medium effect (Cohen's D) was found for this increase. This is likely due to the sheer amount of exposure to texts that students had during this intervention. On average, students were exposed to over 500 texts in seven 25-minute gaming sessions (see Chapter 3). Dourda et al. (2014) also integrated a game into the classroom context and while they did not measure reading comprehension on pre- and post-assessments, they did conduct classroom observations and noted that repeated exposure to large amounts of in-game texts led students to employ a variety of reading strategies, which improved their reading skills over the course of the 8-week project.

The second research question explored the effect of prior gaming experience on learning outcomes. Researchers have suggested that prior gaming experience can help reduce the cognitive load required of a student when playing an educational game (e.g.

Hsu & Wang, 2010). In other words, if a student is more focused on manipulating the basic controls or user interface of a game, they may spend less time engaging with the game in a meaningful way that benefits learning. This study did not find a significant relationship between prior gaming experience and learning outcomes. This is an important finding because we want the use of games in classroom instruction to have a positive impact on all students and not just those who have prior gaming experience. However, it should also be noted that a lack of a significant association between prior gaming experience and learning gains may have been due to the in-class support provided.

The third research question explored the type of support that students need when integrating a digital game into a DLI classroom. Researchers have suggested that learners need some form of support while playing digital games to learn (Cornillie et al., 2012; Peterson, 2011; Rankin et al., 2006) and more recently there has been a more direct call for teacher involvement in L2 learning via digital games (deHaan, 2019; Jones, 2020). The present study addresses these calls by identifying seven types of support that were either requested or given while students played a digital game in the classroom. These supports include quest management, battle strategies, talk around the game, encouragement, technical support, confirmations, and linguistic support. These seven types of support can be broadly parsed into two categories: supporting literate game users and language support. Literate game users are players that "are able to recognize the game's rules and generate strategies to meet the goal of the game" (Hsu & Wang, 2010, p. 407). This first category includes *Quest Management*, *Battle Strategies*, and *Technical*

support. These include helping players follow quest lines, discussions about tactics to be used in battle, and support in learning the game's user interface and basic rules. On one hand, the need for these supports could indicate a lack of gaming experience by the players. In a review on using games for vocabulary learning, Yudintseva (2015) argues that players need scaffolding to learn new game mechanics and cultural norms in-games. On the other hand, for game designers, this need for gaming support could indicate poor game design that did not allow players to learn on their own through in-game scaffolding. However, from a classroom perspective, it provided several opportunities for meaningful discussions around the game. Not captured in this study were the conversations between peers around the game as well. If games are designed to be stand-alone learning tools, in which players do not need support, these opportunities for rich L2 interactions will be lost.

The second category of supports refer to generic L2 supports and includes *encouragement, confirmations, talk around the game, and linguistic supports.* The encouragement support pushed learners to read on their own and focus on important texts. Confirmations allowed for quick reassurances that a player was on a correct path or understood a text. Similarly, linguistic supports allowed players to gain quick access vocabulary knowledge and then to return to the text and in some cases linguistic support allowed for additional discussion around a word. Finally, the talk around the game allowed learners and teachers to use the game as a mediator for discussion (Poole et al., 2019). Teachers were able to point to different parts of the game or the student's character in the game and make a simple comment that sometimes led to a further

discussion. In summary, these supports provided further opportunities for L2 vocabulary, reading, and oral language learning.

Conclusion and Implications

Research on games in L2 contexts has suggested that learners need support to develop their language knowledge and skills. However, few studies have explored what that support might look like in a classroom setting. For digital games to reach their full potential as an educational tool, it is imperative that researchers continue to explore ways to integrate them into the classroom and further how teachers can leverage games to enhance L2 instruction. In the present study, I presented the design of an educational game and supplemental materials that were created to support Chinese vocabulary learning and reading comprehension skills. Then I provided in-depth descriptions of how the game was integrated into the classroom and the role that the instructors provided. After seven gaming sessions over 4 weeks, learners had significant gains in vocabulary knowledge and reading comprehension. In the present study, I do not attempt to parse out the specific role of the game in relation to learning. Instead I argue that it was the integration of the game into the classroom and its ability to provide an environment rich in L2 supports that led to learning gains. The types of support that were provided by the researcher of this study were coded in to seven categories and then more broadly into gaming literacy and linguistic supports. Future research should explore how teachers, who may not be gamers or have insider knowledge of the game, provide support when playing L2 games and how such support leads to learning. Further, studies should explore how positioning games within a curriculum or leveraging in-game content to reinforce

content taught outside of the game affects learning. Finally, this study did not explore the effect of peer-to-peer interactions that occurred during gameplay. Future studies should explore how digital games in the classroom allow for opportunities to engage in meaningful discourse around the game.

Limitations

There are several limitations that are associated with this study. First, there were two primary instructors involved in this study, but support from only one of the instructors were transcribed and analyzed. Future research should explore how different teachers working with other games provide in-class support. Second, only audio data was collected and thus it was difficult to identify specific students associated with each teacher support utterance. Thus, it may be that certain supports were only provided for certain student types. For instance, there were several students that never requested help within the game. Finally, there was no control group in this study, thus the generalizability of the results is limited and it is unclear how effective this game was in comparison to traditional L2 instruction. However, given the goals of this dissertation and the further benefit of being able to collect student data via the game, this was seen as an acceptable limitation.

CHAPTER 3

APPLYING EDUCATIONAL DATA MINING TECHNIQUES TO EXPLORE L2 LEARNING IN A DIGITAL GAME

Abstract

This study explores how educational data mining approaches and data driven explorations can provide insight into how players interact with a game and how those interactions relate to learning. In this chapter, the use of learning analytics and educational data mining approaches in L2 learning research is explored. Then visualizations techniques, classification and regression tree analyses, and cluster analysis are used to explore how in-game indicators, such as battling and use of a glossing tool are associated with learning and affect change. This study found that time on task and use of the glossing tool were the most important variables in determining learning gains. In addition, four subgroups of players were identified based on their gameplay styles. However, no significant differences in learning or affect were found based on the subgroups identified by the cluster analysis.

Introduction

Research investigating the application of digital games for second language (L2) learning and teaching have steadily been on the rise over the last 20 years according to three systematic reviews in that time period (Cornillie et al., 2012; Hung et al., 2018; Poole & Clarke-Midura, 2020). One of the advantages of doing research with digital

games is the ability to collect large amounts of data related to gameplay and learning. The most recent review of digital games being used for L2 learning noted that only 28.5% of the studies in the review took advantage of log data collected while participants played the game (Poole & Clarke-Midura, 2020). Many of these studies used log data as a means to retrieve answers on in-game quizzes (e.g., Erhel & Jamet, 2016; Ming et al., 2013) or to collect chat dialogue to be analyzed manually (e.g., Bytheway, 2014; Rama et al., 2012). While these applications of log data are beneficial to researchers, this data could be further explored using educational data mining (EDM) techniques.

EDM is the process of making sense of large data sets with multiple variables (Baker & Inventado, 2014) and is often compared to learning analytics (LA) because both EDM and LA are concerned with using big data sets to investigate, inform, and improve education and learning. However, Siemens and Baker (2012) highlight a few key differences between the two research communities. They argue that EDM is more focused on using automated approaches to discovering patterns within data and then using human judgement to decide what the patterns mean, while LA uses human judgement to define patterns and then automated approaches to apply the human-defined patterns. Further, they argue that EDM approaches are often used to inform the development of automated learning systems, such as for personalized learning environments. While LA models are used to inform practice and empower instructors. Finally, EDM approaches tend to reduce the data to make sense of it, while LA tries to understand the "systems as wholes" (Siemens & Baker, 2012, p. 253).

In a recent call for a larger focus on EDM and LA approaches in L2 learning,

Reinders (2018) points out that big data collected from learning management systems is not only valuable for L2 researchers, but could also inform L2 teacher practices in the classroom. This commentary identified teachers as an audience and provided several examples illustrating how big data could be used to identify struggling students or improve instruction. However, as a recent review on EDM in an L2 context illustrated, few studies have explored data collected in K-12 settings thus limiting the opportunities for educators at the primary and secondary level (Bravo-Agapito et al., 2020).

While clearly there are differences between EDM and LA, these differences are often defined in terms of tendencies, suggesting that there is considerable overlap. Similarly, for the present study, automated EDM approaches, with reductionist goals, are applied to data derived from a digital game, but they are done so with the intent of informing teacher practices with digital games in the classroom and future game designers. Thus, the present study explores data collected from a digital game that was integrated into in an elementary classroom context to promote vocabulary acquisition and reading comprehension skills (see Chapter 2). Specifically, the present study investigates how in-game behaviors are associated with gameplay styles and learning.

In the sections that follow, I first provide a review of the literature, including a brief overview of approaches associated with EDM, followed by a review of past studies that have implemented EDM approaches in an L2 context. I then discuss past research that has taken advantage of data collected via digital games and the approaches used with such data. Then I introduce the methods, including a brief description of the game used in this research. In this section, I review the EDM approaches that were applied in this

study. The results illustrate how EDM approaches applied to a digital game can provide insight into how the game promotes learning and further how students play the game in unique ways. Implications for these findings are explored in the discussion and conclusion sections.

Literature Review

Educational Data Mining

Educational data mining applies data mining approaches and techniques to data derived from an educational context. Data mining has been defined as the process of discovering hidden patterns in big data sets (Fayyad et al., 1996). Further, EDM is concerned with extracting and processing raw data that is often unstructured from unique environments to make sense of a learning environment (Romero & Ventura, 2010). To accomplish these goals, Siemens and Baker (2012) identified five general categories of approaches applied in EDM: prediction, clustering, relationship mining, discovery with models, and distillation of data for human judgement. Prediction is concerned with identifying a variable within a dataset using a combination of other variables. Clustering attempts to find subgroups within a dataset by grouping cases together based on common data points. Relationship mining explores types of relationships (e.g., associations, sequential patterns, causal) between variables within a data set. Romero and Ventura point out that these first three approaches are common techniques used in data mining, however the last two (discovery with models and distillation of data) are more uniquely EDM approaches. In discovery with models, a model is developed from one of the first

three approaches and then is used as part of another analysis. Finally, distillation of data for human judgement applies visualization techniques to data to allow for human inferences and interpretation. In the following section, research that has applied these approaches to an L2 context are reviewed.

Educational Data Mining in L2 Research

Godwin-Jones (2017) points out that using data derived from digital environments is not new for L2 researchers, it is the sheer size of datasets that are being produced that is novel. He argues that this is largely due to increased use of Learning Management Systems (LMS), personalized learning environments, and the creation of other digital software for learning. In a study exploring data from an LMS, Foung (2019) used classification decision trees to develop a model to predict English learners who were at-risk of dropping out of a course. The author created models with approximately 90% accuracy in detecting at-risk students, which was then argued to benefit teachers and administrators by providing real-time analytics on learners. Further, without these data, teachers may not be able to identify these at-risk students until it is too late. In a study that explored student behavioral patterns in a digital reading environment, H. Lee et al., (2019) used clustering approaches to explore unique ways that students interacted with a textual glossing system. In an earlier study with the same data, the authors found that students using the glossing system reported the highest post-test scores on average compared to two other conditions (H. Lee et al., 2017). However, when using cluster analysis to explore potential hidden patterns in a follow-up study, the authors found that L2 proficiency was a strong determinant in the effectiveness of the glossing system.

Finally, in a study that explored how students collaborated in Google Docs, Yim et al. (2017) used visualization and text mining techniques to identify four styles of collaborative writing. These styles of collaborative writing were then explored to determine which collaboration style led to higher quality writing. The visualization and text mining techniques used in this study allowed the researchers to explore and evaluate writing processes in an unobtrusive way. In the following section, I review the research on analyzing data from digital gaming environments, including two studies that applied EDM approaches to data originating from a digital game.

L2 Gaming Research and Log Data

Data collected in digital games in the past 20 years has been used in a variety of ways. A majority of the studies examined chat logs to investigate language used within a gaming environment (e.g., Bytheway, 2014; Peterson, 2011, 2012; Rama et al., 2012; Rankin et al., 2006; Vosburg, 2017; D. Zheng et al., 2012, 2015). Some studies have designed quiz-like tasks into games and used the log data to export scores for analysis (Erhel & Jamet, 2016; Cornillie et al., 2012; Ming et al., 2013).

Other researchers have attempted to use in-game variables, or data, as indicators for the effect that playing a digital game has on learning outcomes and L2 proficiency. For example, in a study exploring the effect of time spent on reading a text, Collentine (2011) found that time spent on task was significantly associated with outcomes on a writing assessment after playing a digital game designed for second language learners of Spanish. In a study investigating L2 learning in Everquest II, Rankin et al. (2006) found that a variable counting the number of messages produced was significantly associated

with more advanced English speakers. In other words, more advanced speakers generated more chat messages than intermediate and novice speakers. Using in-game actions as predictors can be one method of illustrating how a game promotes learning or predicts proficiency.

A few studies have applied EDM techniques to digital games used in a L2 context to explore how gameplay behavior is related to affect and learning. For example, Hwang et al. (2017) used a progressive sequential analysis to compare gameplay sequences by EFL learners with high and low L2 learning anxiety. They found that learners with higher levels of anxiety engaged in more complex forms of learning than less anxious students. They further found that high anxiety learners reviewed more vocabulary and acquired more relevant knowledge before completing tasks. Hsiao et al. (2017) explored behavioral patterns around players interacting with new vocabulary words in a virtual environment. Players were told to learn 30 words while interacting with the environment. Through data visualization techniques, the authors found differences in learning strategies between high- and low-achieving learners. They conclude that low-achieving learners tended to click on vocabulary words randomly or use the nearest neighbor approach in which they simply clicked on vocabulary words that were close to each other. Whereas, high-achievers tended to use a strategy to cluster similar vocabulary words together to facilitate learning. These EDM approaches allow researchers to see how games are actually being played and further how such gameplay styles are potentially associated with changes in affect or learning.

The present study looks to build on past research implementing data mining

techniques in L2 contexts that use digital games. Specifically, the present exploratory study uses visualization, decision trees, and clustering approaches to explore how in-game variables are associated with learning gains, and further, how playstyles may be associated with change in learning and/or affect. The present study was guided by the following research questions.

1. What, if any, in-game variables are associated with change in vocabulary knowledge and reading comprehension, as measured by the pre- and post-assessments?

2. Are in-game playstyles associated with affective factors and/or vocabulary knowledge and reading comprehension gains, as measured by the pre- and post- assessments?

Methods

Materials

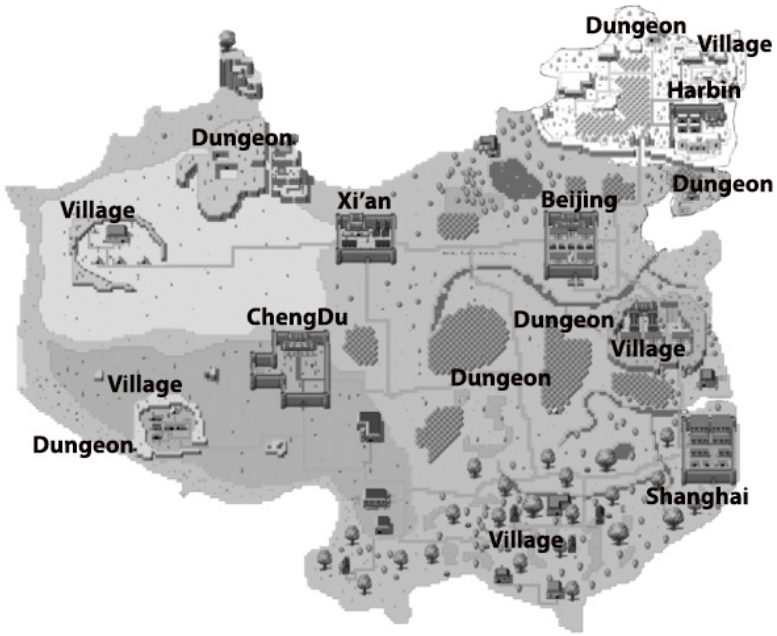
The digital game being used in the present study (see Chapter 2) was designed to promote vocabulary learning via an in-game glossing system (Poole & Sung, 2016) and reading comprehension via series of quests and in-game tasks that are driven by text-based dialogue. In this game, students take on the role of an adventurer who sets out on a quest to save the last dragon in China. As they carry out quests, students meet non-player characters (NPC) who provide information, present quests/tasks, and direct students towards the last dragon (see Figure 3.1). The game world consists of five major Chinese cities and several fictional villages and dungeons placed in proximity to the cities. The game world was designed to resemble the shape of China with cities located in their approximate real-world locations (see Figure 3.2).

Figure 3.1

Example Quest

Figure 3.2

Over World Map of Legend of the Dragon



Participants and Setting

Data were collected in two sixth-grade elementary Chinese DLI classroom settings. The two classrooms had 19 and 21 students respectively. Stratified randomization with gender as a factor was used to assign learners to each classroom at the beginning of the school year. Students ranged in age between 10 and 12 years old (mean = 11.05). Four students (1 in Class A, 3 in Class B) did not sign the consent form. In addition, two students in Class B did not take the pre-assessments, and one student in both classes did not take the post-assessments. Thus, although I had consent forms for 36 students, there was only complete data from 32 students (see Table 3.1).

Table 3.1

Students Included in Data Analysis

Class	$N_{students}$	Survey data	Log data
A	20 (10F, 10M)	20 Pre, 19 Post	20
B	16 (9F, 7M)	14 Pre, 15 Post	16
Total	36 (18F, 16M)	34 Pre, 34 Post	36

Note. Only 32 complete cases.

Procedures

On the Friday prior to gameplay, the pre-assessment was administered with a randomized version of the vocabulary and reading comprehension assessments via paper. In the sixth-grade classroom, there is a 1:1 computer to student ratio. However, *RPG Maker MV* games are not compatible with Chromebooks. Thus, the researcher brought in 11 MacBook laptops to play the game. In order to maintain a 1:1 ratio, the class was divided into four groups of five students each, with a mixture of high and low proficiency

learners as identified by the class instructor. Two of the groups played the game while the other two groups completed the workbook. After 25 minutes, these groups switched.

On the first day of gameplay, the participants were given a brief tutorial on how to play the game via a whole class demonstration. The classroom teacher then illustrated how the game related to their current studies by telling the students that vocabulary in the game was primarily comprised of review words and that the game was similar to their current in-class readings in that they were reading about students who were traveling to China. Similarly, in the game the students would be going to China and could explore the same places that the characters in their books were visiting (e.g., Beijing, Xi'an, Chengdu).

Participants played the game for 50 minutes (two 25-minute sessions) and completed supplemental materials for 50 minutes (two 25-minute sessions) per week over the course of a 4-week period (once on Monday and once on Wednesday). Participants played the game for approximately 3 hours over the four weeks. Students were administered the post-assessment two days following the final game play session. Figure 3 below illustrates the timeline for the study.

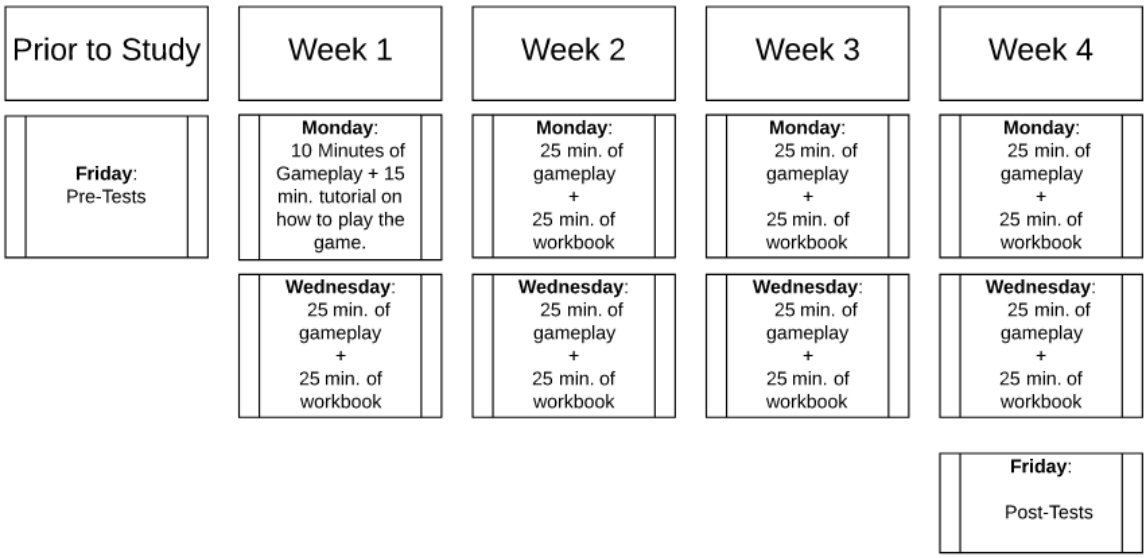
Data Collection

Measures

Data for the present study came from four sources: pre- and post-affect surveys, pre- and post-vocabulary assessments, a post-reading comprehension assessment, and log files that captured in-game actions, texts read, and choices made.

Figure 3.3

Research Schedule



Affective survey. The pre-affect survey consists of background information (e.g., name, age, gender, and language spoken at home), two questions on gaming background, and seven items regarding Chinese reading anxiety using an 8-point Likert scale. Reading anxiety was added to the affect survey as past research indicates that anxiety is often a factor in reading comprehension for novice learners of Chinese as a second language (Zhao et al., 2013). Further, L2 anxiety was found to influence in-game behaviors in past research (Hwang et al., 2017). The items for Chinese reading anxiety were adapted from Saito et al. (1999). Saito et al. originally had 20 items to assess foreign language reading anxiety and validated the items with a Cronbach's alpha of .93 for the scale. The number of items was reduced to seven items to make the survey as brief as possible. The post survey consists of the same seven items concerning Chinese reading anxiety, and ten items about the gaming experience adapted from De Grove et al. (2010). De Grove et al.

reported a Cronbach's alpha ranging between .63 and .88 for the gaming experience scale and includes the following constructs: positive affect, negative affect, challenge, frustration, and immersion. However, for this study only positive affect, which explores student enjoyment, and reading anxiety measures were used.

Vocabulary assessment. The pre- and post-vocabulary assessment consisted of 45 words that could be found in the game. Learners were prompted to enter the pinyin and the English for each vocabulary word presented in character form. The vocabulary assessments can be seen in Appendix B. These assessments were scored by awarding 1 point for correct pinyin, and 1 point for correct English definitions. Further, half points were awarded for partial answers. For example, if a learner correctly identified the correct pinyin or English for one of the characters but not both half points were awarded. Awarding partial points for vocabulary knowledge was viewed as valuable given past research that has noted the non-linear trajectory, and partial accumulation of vocabulary knowledge that occurs through incidental learning while reading L2 texts.

Reading Comprehension Assessment. The external reading comprehension assessment was adapted from the Chinese YCT test (http://english.hanban.org/node_8001.htm) which is an official Chinese proficiency assessment developed by the Confucius Institute and used regularly in the Utah Chinese DLI program. The reading comprehension assessment consisted of ten items. See Appendix A for the external reading assessments. Although the format of the reading comprehension assessment was adapted to reflect the Chinese YCT test, the content was adapted to reflect text that the learners might see in the game. To reduce the priming effect on the comprehension

assessments, items were randomized in both the pre- and post-assessments. Further, although the sentence structures remained the same, the content, and thus the answers changed from pre- to post-assessments. Finally, gain scores for both vocabulary learning and reading comprehension were calculated by subtracting pre-scores from post-scores.

Log Data

Five key variables were collected from the in-game log files: (1) *Text exposure* counted the total number of texts that a player was exposed to while playing the game. (2) *Battles* counted the total number of battles that player engaged in. (3) *Menu-On* counted the total number of times that a player accessed the menu to either look at a skill, card or item that was acquired. (4) *Average time reading* was a variable that tracked how long students spent reading each text in seconds on average. (5) *Look up* tracked how many times the players used the glossing system in the game to look up an unknown word. Table 3.2 provides an overview of data that were collected and used in this study.

Data Analysis

In the present study, the relationship between in-game variables, affective factors, and learning are explored with educational data mining (EDM) approaches. EDM involves the use of algorithms from statistical, machine learning, and/or data mining fields to explore educational data (Romero & Ventura, 2010). In the following section, I first describe the process for preparing log data. Then I describe how classification and regression tree (CART) analysis along with clustering analyses were used to explore the two research questions in this study.

Table 3.2

Data Collected for this Study

Data Source		Time	Description
Affect	Positive affect	Post	This measures student enjoyment while playing the game.
	Reading anxiety	Pre- and Post-	This measures student anxiety associated with reading Chinese texts.
Knowledge/skills	Vocabulary assessment	Pre- and Post-	This measures student knowledge on 45 vocabulary words that are found in the game.
	Reading comprehension assessment	Pre- and Post-	There are 10 reading comprehension questions that use vocabulary and sentence structures from the game.
In-game logfiles	Battles	In-game	This variable is important because it provides insight into whether players were battling more or following quests more.
	Text exposure	In-game	The number of texts that a player is exposed to provides insight into how far the player progressed in the game.
	Menu-on	In-game	This a raw account of how many times the menu was accessed during gameplay.
	Time spent reading	In-game	This was a measure of time spent reading text in seconds. Time was rounded to the nearest second. This variable can provide insight to how carefully a text was read on average.
	Look up (vocabulary)	In-game	This counts the number of times a word was looked up in the text and is important to determine if a player was focused on understanding a text.

Log Data Processing

The source of in-game data came from the log files that capture players' in-game actions, texts read, and choices made. All players' movements and choices along with a timestamp and were saved to a JSON file after each session's gameplay and then automatically uploaded to a Mongo Database each time a player saved the game. After

the study was completed, all JSON files were pulled from the Mongo Database and stored into one master JSON file (see Appendix E for sample Log Data in JSON format). This file was then wrangled into a single data frame using primarily *tidyverse* (Wickham, 2017). To extract this data and wrangle it into a data frame, the knowledge discovery in databases (KDD) process was used. Knowledge discovery refers to the "nontrivial extraction of implicit, previously unknown, and potentially useful information from data" (Frawley et al., 1992, p. 58). This is an iterative process that involves five broad phases: selection, pre-processing, transformation, data mining, and finally interpretation/ evaluation (Fayyad et al. 1996).

The selection phase refers to the process of identifying data of interest based on the target domain. In the case of this study, a JavaScript plugin was adapted and added by the researcher to the digital game to track in-game actions with timestamps. The pre-processing phase refers to data cleaning and removing missing or irrelevant items from the dataset. In this study, the pre-processing phase was primarily focused on wrangling the data into a wide format rather than a long format. In other words, when the data was collected, every action was stored in a separate row and thus responses to questions were spread across three rows making it difficult to associate a text with a student response. Further, there were several actions captured, such as every movement change and moving between maps that needed to be removed. Before these actions were removed, there were approximately 520,812 cases. It was at this point that the time variable had to be captured, because after removing the time stamps associated with moving it would be impossible to know if time between texts was a result of moving or reading the text. After

capturing the time variables for text being read, these movement variables were removed, and then the data was collapsed into a wide format to accommodate as much information around the texts as possible. After this stage there were 19,781 cases remaining. This data set was then summarized by learner and used for the analyses in this paper. The transformation phase refers to the process of transforming data into variables that are appropriate for analysis. This included calculating a time variable for each text by subtracting the time from the action that preceded a text event from the start of the text event, as mentioned above. This phase also included calculating the length of texts read and converting the glossing tool event into a binary variable. Finally, the last two phases, data mining phase and interpretation phases are concerned with choosing data analysis techniques and then interpreting them. These will be discussed further in the analysis section below.

Classification and Regression Tree Analysis

The first research question explores which in-game variables are important indicators for L2 learning. The analysis used to answer this first question is a classification and regression tree (CART) analysis. The CART analysis is an approach often used in data mining or machine learning. It involves a type of supervised learning because the outcome variable is first identified and then a decision tree is created to select the most relevant variables. This analysis uses an algorithm that chooses variables that will reduce the sum of squared errors in a regression model each time a partition is made (a variable is selected). When the outcome variable is a discrete, a categorical variable, the decision tree is called a classification tree and when a continuous variable is used it is

called a regression tree. When used for predictive purposes, these analyses often apply a random forest approach in which several decision trees are created to improve the accuracy of predictions (M. Hong, 2018). Variables that have little or no effect on the outcome variable will not be selected by the algorithm in the analysis. This makes the CART analysis ideal for data exploration that seeks to understand which variables are most important indicators of a target variable (Kazemitabar et al., 2017; Song & Lu, 2015). In the present study, CART analysis will be used to identify variables that are deemed important to learning within the game. For these analyses, learning is defined as gains in vocabulary knowledge and reading comprehension scores as measured by the pre- and post-assessments.

CART analysis is an ideal approach for the data collected in the present study for a few reasons. First, data within the game is comprised of both categorical and continuous variables. Second, variables from in-game data are on very different scales and are likely not normally distributed. Finally, given the autonomy that students are allotted within the game and the vast number of ways through which the game can be played, there is a lot of missing data. In other words, some students may have been exposed to 700 passages and read them at quick speeds, while other students only read approximately 300 passages but more carefully. CART analysis can handle both categorical and continuous data; it is also good at dealing with outliers and missing data making it ideal for log data which can be messy (M. Hong, 2018; Mendez et al., 2008; Song & Lu, 2015).

The first step in setting up a CART analysis is to identify the outcome variable. In

the present study, the two outcome variables were vocabulary learning gains and reading comprehension gains. The next step is to enter variables that are seen as relevant to the outcome variable. Table 3.3 explains the five variables, collected via log data, that were entered into the model in the present study.

Table 3.3

Variables used in CART Analysis

Variable	Description
Text	The total number of texts read by a student.
Menu On	The total number of times a student accessed the menu.
Battle Start	The total number of times a student engaged in battle.
AvTime	The average amount of time a learner spent reading a passage in the game. Measured in seconds.
LookUpCount	The total number of times that a learner looked up a word while playing.

The algorithm then cycles through the predictors and chooses the predictor that will best parse the selected outcome, and the cutoff point for that variable that best reduces the sum of squared errors. The goal of this analysis is to create the most similar groups according to the outcome variable. This process is repeated over and over again until the minimum sum of squared errors is reached. This process is called recursive partitioning. The cutoff point is called the complexity parameter and controls how large the tree grows (Song & Lu, 2015). A cutoff point of .01 is often the default. This would dictate that once the splitting of the outcome variables fails to reduce the sum of squared errors by at least .01, then the model stops. The goal is to find a model that is not too complex (i.e., overfit) nor too simple. To determine the ideal complexity parameter, a

series of trees are built from the smallest tree to the largest tree, then the cross-validated error for each tree is examined (Therneau & Atkinson, 1997). Once the cross-validated error begins to increase (instead of decrease) as a result of adding a split, the decision tree should stop building (Williams, 2011). Thus, the goal is to identify the lowest cross-validated error and the complexity parameter that is associated with the lowest cross-validated error. In the present study, a complexity parameter of 0.06 was chosen for the vocabulary learning gains CART analysis, based on examining cross-validated errors from the default model. A complexity parameter of 0.11 was used for the reading comprehension gains CART analysis by examining the cross-validated errors from the default model. All analyses were conducted in R, using the Rpart package (Therneau & Atkinson, 2019). This analysis will be discussed in further detail below in the results section, which is where the CART analysis visuals can be found

Cluster Analysis

The second research question explores in-game behaviors and their potential relationship to learning and proficiency in an unsupervised cluster analysis. Cluster analysis can be used for data exploration or to test hypothesis about data structures (Huberty et al., 2005). The present study applies cluster analysis as an exploratory approach to student gameplay data. Cluster analysis is useful for exploring groups within a set of participants who share similar characteristics (Warschauer et al., 2019). Further, cluster analysis is useful to explore potential patterns or groupings that are not easily seen from other analytic approaches. In the present study, participants who have similar gameplay styles are explored. More specifically, this approach is used to identify distinct

in-game behaviors, such as time spent on text, the amount of battles engaged in, and the use of the glossing tool, that may be associated with vocabulary gains and affective factors related to L2 learning, namely reading anxiety and game enjoyment.

In the present study, I employ a K-means cluster analysis. All variables were normalized before beginning the analysis to prevent any inflated differences due to differences in scales. Formann (1984) suggests that the minimal sample size for cluster analyses is no less than 2^k, in which k refers to the number of variables included in the analysis. The preferred amount would be 5*2^k. Thus, for this project, with only 36 students who have complete gameplay log data, five variables (2^5=32) is the maximum number of variables that should be used, with less variables being more ideal. With this in mind, variable selection must be strategic. Four of the variables from Table 3 that were shown to be important in the CART analysis were added to the cluster analysis including: the number of look ups, average time spent on reading the text, total number of battles engaged in, and the total of number of texts read.

The next decision to make when doing a cluster analysis is the algorithm to use. K-means algorithm is a type of centroid clustering that searches for center points of clusters in a way that minimizes the distance of each member of each cluster. K-means clustering also involves indicating how distance is measured (Attewell, Monaghan, & Kwong, 2015). When using a k-means approach, the number of clusters k must first be identified. One approach is to simply try a different number of defined clusters and then using a cluster plot and a silhouette plot, compare which number of clusters bests divides the data into meaningful groups. However, the elbow method can also be run to

determine the optimal number of clusters. In the present study, Figure 3.4 (left) illustrates the total sum of squares that is accounted for by each number of clusters. The point at which the line begins to level out is when adding more clusters no longer increases the sum of squares accounted for by the analysis. Thus, in this figure, either 3 or 4 clusters appears to be ideal. Figure 3.4 (right) provides a similar analysis, but uses the average silhouette width, which is the measure of how similar an item is to its cluster. Thus, the higher the average similarity score for each item in a cluster the better. After four clusters, the average silhouette score begins to drop suggesting that a 4-cluster solution is the most ideal.

To further validate the use of 4 clusters in this analysis, Figure 3.5 shows the cluster silhouette plot. The silhouette analysis can be used to explore how far away clusters are to neighboring clusters and has a range from +1 to -1. Each bar represents a

Figure 3.4

Elbow Method and Silhouette Width

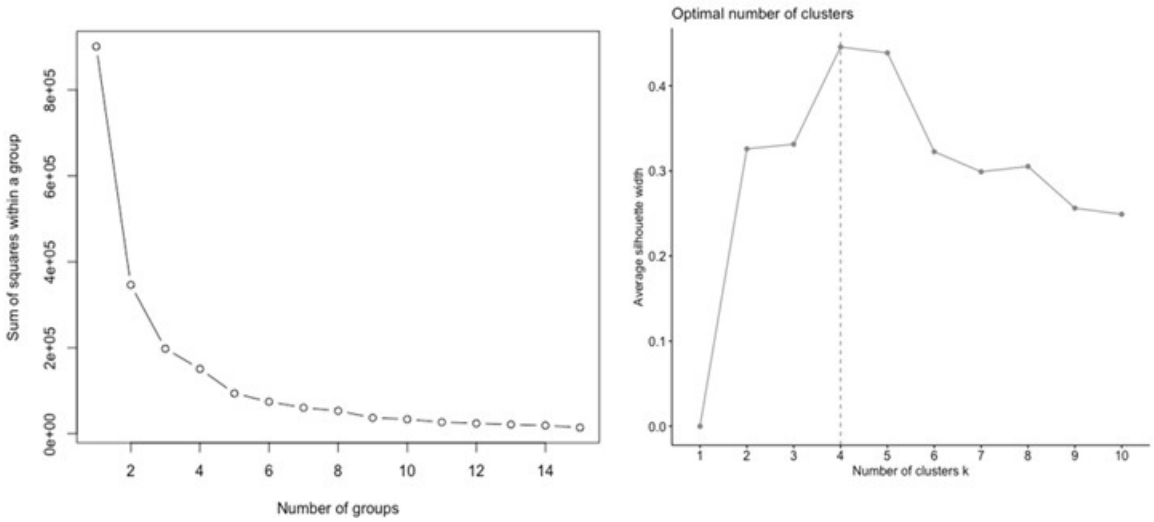
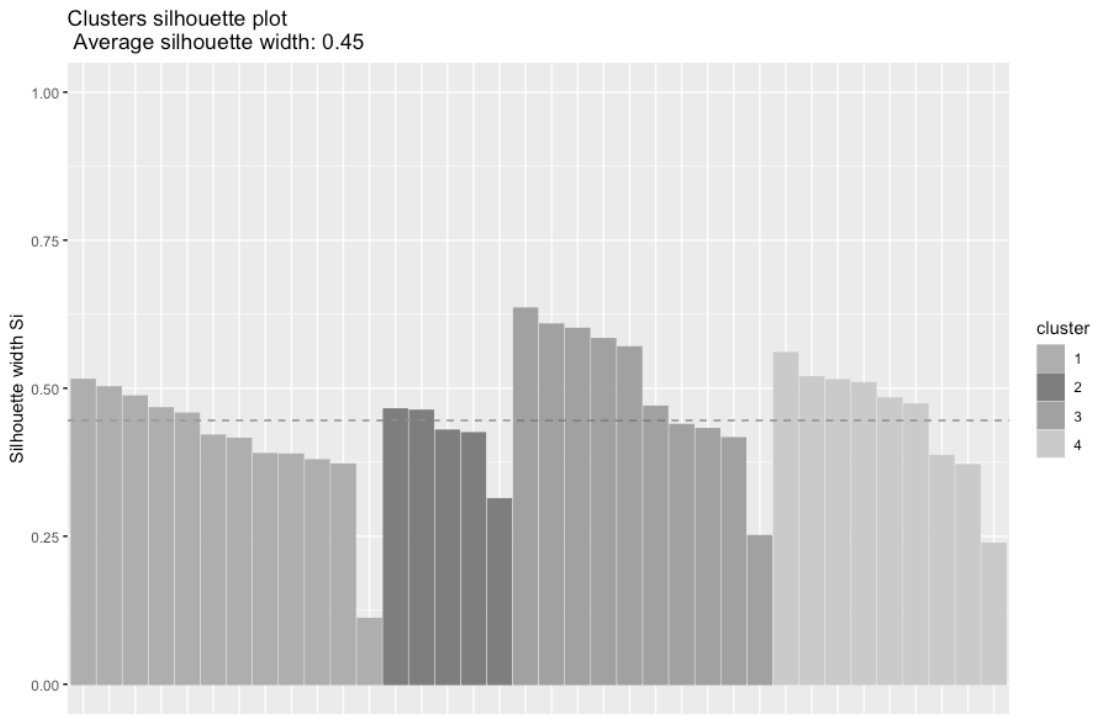


Figure 3.5

Silhouette Plot



student's similarity to their cluster. The dotted line shows the average silhouette score, while any values that are near or below 0 are thought to be students who are close to another cluster. Ideally, none of the clusters will have all of the individual silhouette scores below the average silhouette score, which is the case in Figure 3.5. In this plot, no students were identified as being misplaced, though cluster 1 has one student who is close to 0. In addition, there is an average silhouette width of 0.45, which is the average similarity for all items belonging to a cluster (Rousseeuw, 1987). Thus, this measure can be used to compare clusters. The plot with 3 clusters had 5 misplaced students and an average silhouette width of .26 suggesting that four clusters is better for this analysis (Kassambara, 2017).

After the clusters were identified, a unique identifier was assigned to each cluster and descriptive statistics were run for each group to identify what makes them unique. Further, linear regressions were used to explore if these groups differed significantly in terms of change in learning or affect. These findings will be discussed in the results section.

Results

In the present study, EDM techniques were applied to explore how gameplay relates to learning gains, anxiety, and game enjoyment. To provide context for the results, a table with the learning gains and student affective scores will first be presented. Then two visualizations that illustrate student exposure to vocabulary and average length of texts will be used to provide an overview of student opportunities to engage with in-game dialogue. Finally, a correlation matrix will be used to show how in-game actions are correlated with learning gains.

Learning and Affective Outcomes

Results of the pre- and post-assessments indicate that students who participated in the project had significant gains in vocabulary and reading comprehension. On average, students reported an 9.64 increase in vocabulary knowledge and 1.31 increase in reading comprehension scores (see Chapter 2 for more details). The positive experience construct, a measure of students' enjoyment of the game, found that, on average, students agreed that the game was fun ($M = 6.82$, $SD = 1.27$). Learners on average reported a 0.21 decrease in reading anxiety after the intervention, but this decrease was not significant.

Table 3.4 provides summary statistics, *t*-test results, effect size calculations, and Cronbach's alphas of the measures collected both pre- and post-intervention.

Table 3.4

Measures Used in Analysis

	Pre				Post					
Measures	*M*	Med	*SD*	*A*	*M*	Med	*SD*	*A*	*t* test	Cohen's D
Vocabulary (*n* = 32)	18.31	16.00	14.38	.95	27.95	23.50	17.97	.96	9.99***	0.45
Reading comprehension (*n* = 32)	3.41	3.00	2.56	.74	4.72	4.50	2.43	.60	3.22**	0.54
Positive experience (*n* = 34)	NA	NA	NA		6.82	7.0	1.27	.74	NA	NA
Reading anxiety (*n* = 32)	4.27	4.21	1.27	.65	4.06	4.29	1.28	.72	1.31	NA

Note. Vocabulary Scale from 0-90. Reading scale 1-10. Both effect sizes are medium. Reading anxiety and positive experience are on an 8-point Likert scale. There is no effect size for reading anxiety because the difference was not significant.

** $p < .01$.

*** $p < .001$.

Overview of Text, Vocabulary and In-Game Actions

The goal of the present study is to employ data mining approaches to explore if gameplay behavior or styles are associated with L2 proficiency or learning outcomes. Table 3.5 provides summary statistics of the in-game actions and time spent on text. For the total data set, all students read 19,781 texts with each student averaging approximately 549.5 texts over seven gaming sessions. The average text length was 9.53 characters. On average, students accessed the menu 52 times during gameplay, engaged in battle 19 times, and looked up 37 words.

Table 3.5

Counts of In-Game Actions

Variable	Total	Mean	Min	Max	*SD*
Text	19,781	549.5	190	903	156.27
Menu On	1907	52.77	27	98	18.68
Battle Start	686	19.11	1	55	10.03
AvTime (in seconds)	NA	2.83	1.66	4.57	0.78
LookUpCount	1,295	37.27	0	137	35.46

In addition to providing an overview of in-game actions, it's also important to further explore what vocabulary words the learners were exposed to and consequently what vocabulary words were learned. Figure 3.6 shows each of the vocabulary words that appeared on the pre- and post-vocabulary assessment.

Each of the black and gray dots represent individual students (girls and boys respectively). Dots on the left (0.0) represent students who either knew the word on the pre-assessment or who did not know the word and failed to learn it after the intervention. Dots on the right (1.0) represent those who learned at least partial knowledge of a word after the intervention. It's important to note that some words (e.g., boring) only have one or two dots. This is because these words were only encountered by one or two students. Open-world games allow for students to choose their path and how they play the game. However, because students will explore different parts of the game, and thus, be exposed to different dialogue, all students will be exposed to different vocabulary and different amounts of vocabulary when they play the game.

The next visualization (see Figure 3.7) shows the number of texts and lengths of text that each student was exposed to. While the frequencies differ slightly, every student

Figure 3.6

Vocabulary Learned by Word

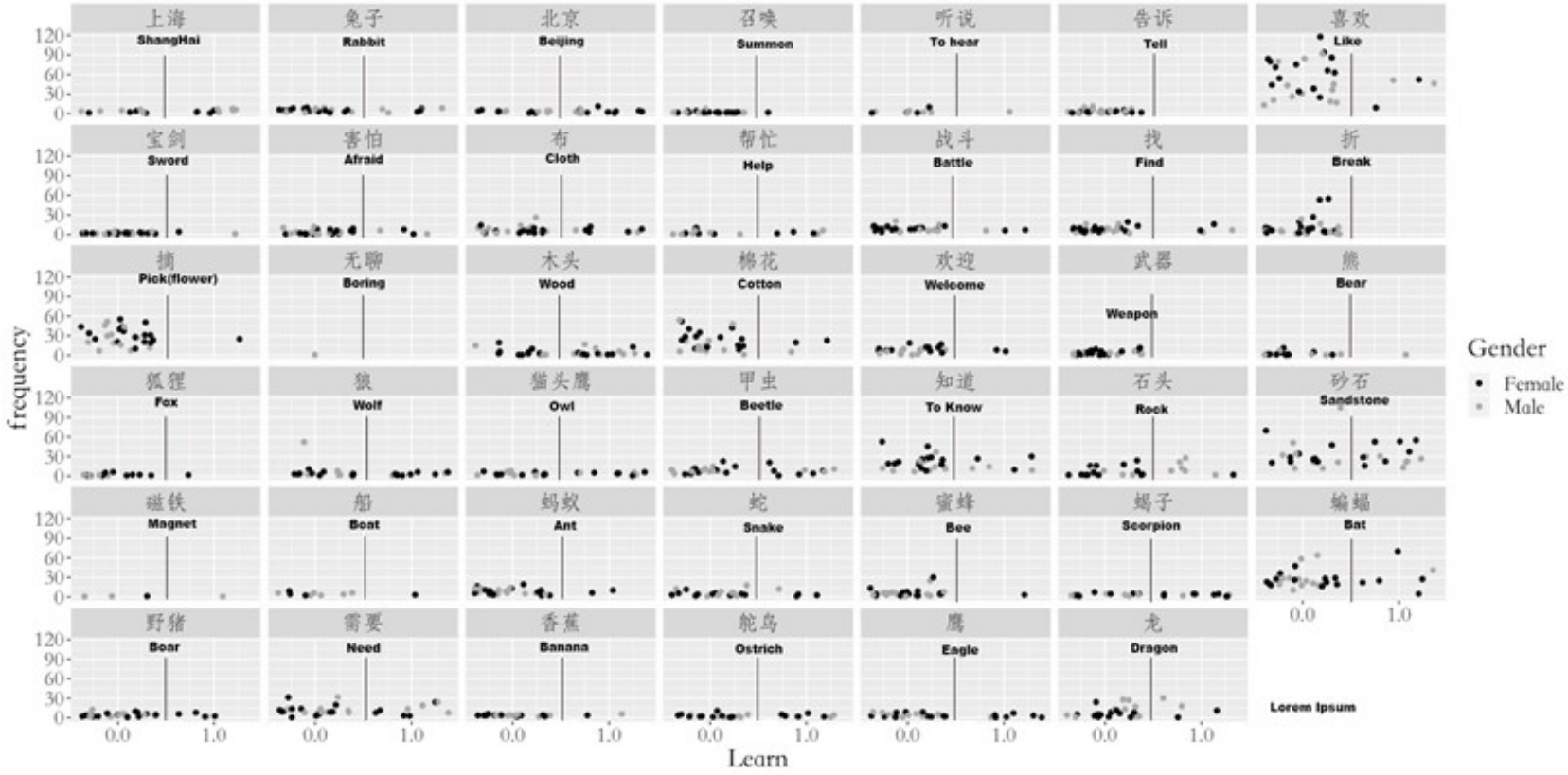
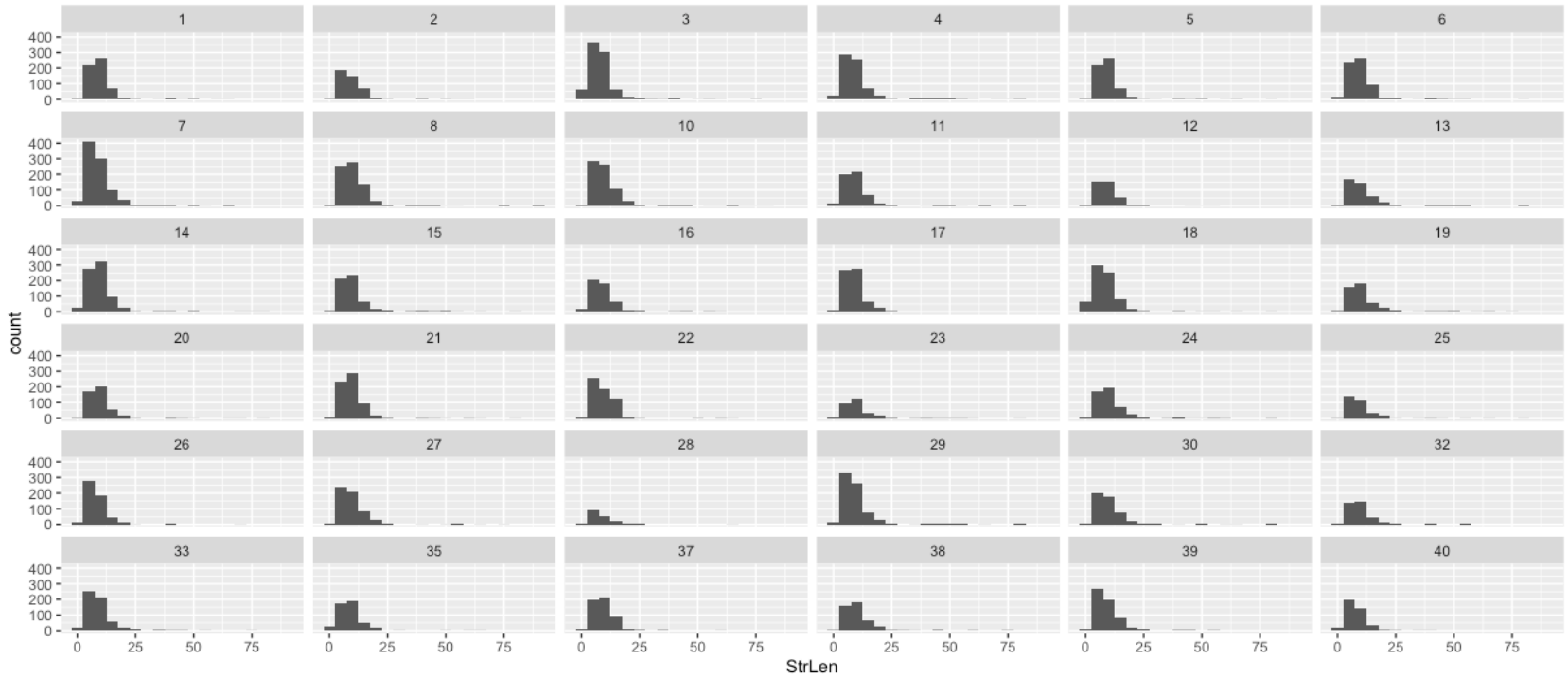


Figure 3.7

Text Length and Frequency by Student



was exposed to a majority of texts that range between 5 and 25 characters long. There were a few students who were exposed to texts in the 30+ character range, but these texts were not frequent occasions.

Finally, a correlation matrix was created to illustrate how in-game actions were correlated with each other and learning gains associated with gameplay. Figure 3.8 shows that none of the in-game measures are directly correlated with learning gains found on the pre- and post-assessment. However, the number of texts read is significantly correlated with battles engaged in and average reading time. In other words, students who engage in more battles are associated with higher exposure to texts. Further, students that read more texts are associated with lower average reading times. This overview was meant to provide some context for the results of the two research questions.

Research Questions

Research question 1 asked, "*What, if any, in-game variables are associated with change in vocabulary knowledge and reading comprehension, as measured by the pre-post assessments?"* In order to answer this research question, I first used CART analysis to explore variables that were potentially important for predicting reading gains. Figure 3.9 shows the results of this analysis. The first node shows that for 31 students, there was an average reading gain of 1.1 points per student. The first variable that splits how students performed on the reading gains assessment is *Menu On*. This variable determines how many times a player opened their menu. Players who did this more than 30 times (n = 28) had an average reading gain of 1.4, compared to those who accessed the menu less than 30 times (n = 3) and had an average 1.7 decrease in reading scores.

Figure 3.8

Correlation Matrix of Learning Gains and In-Game

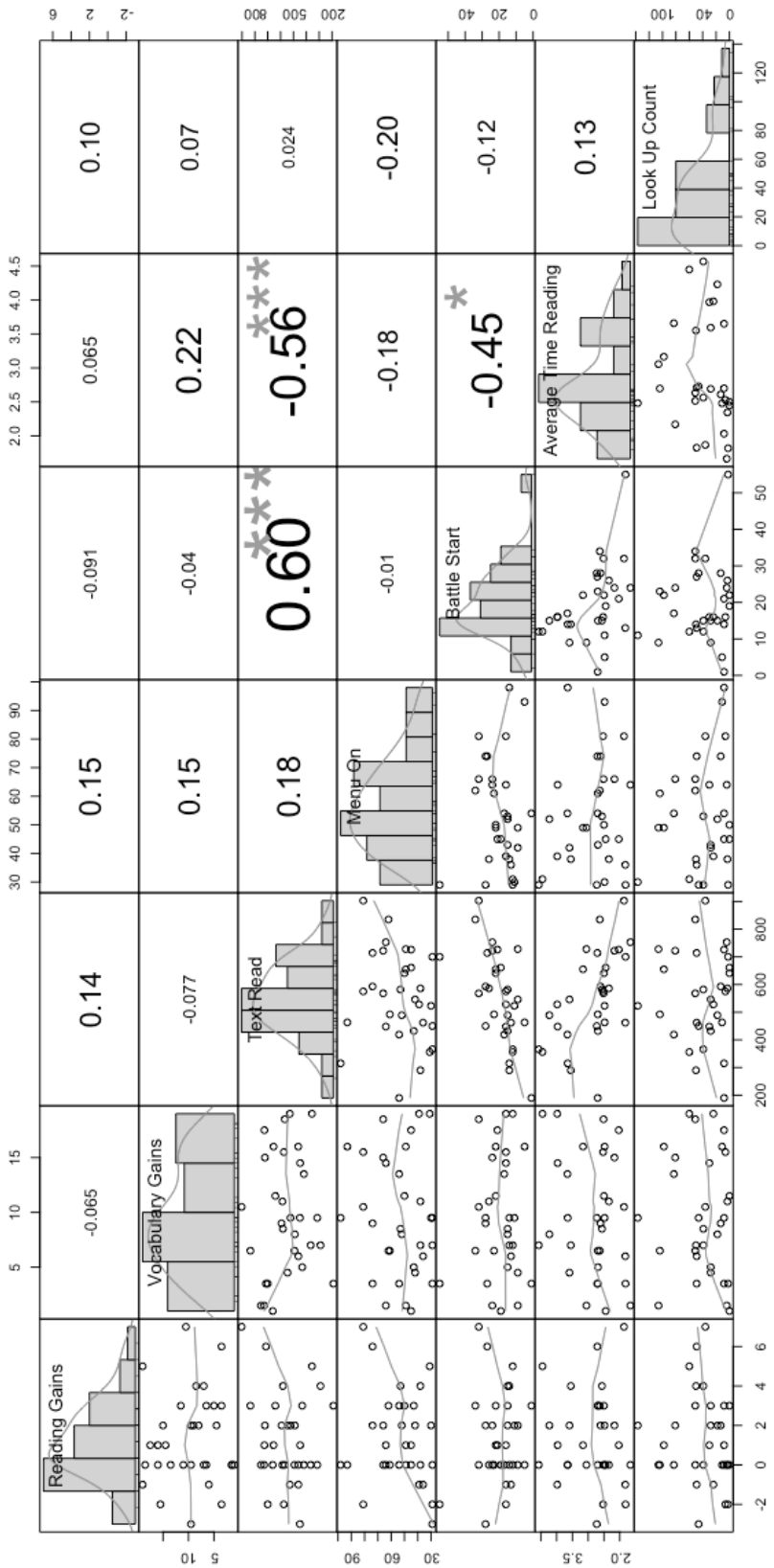
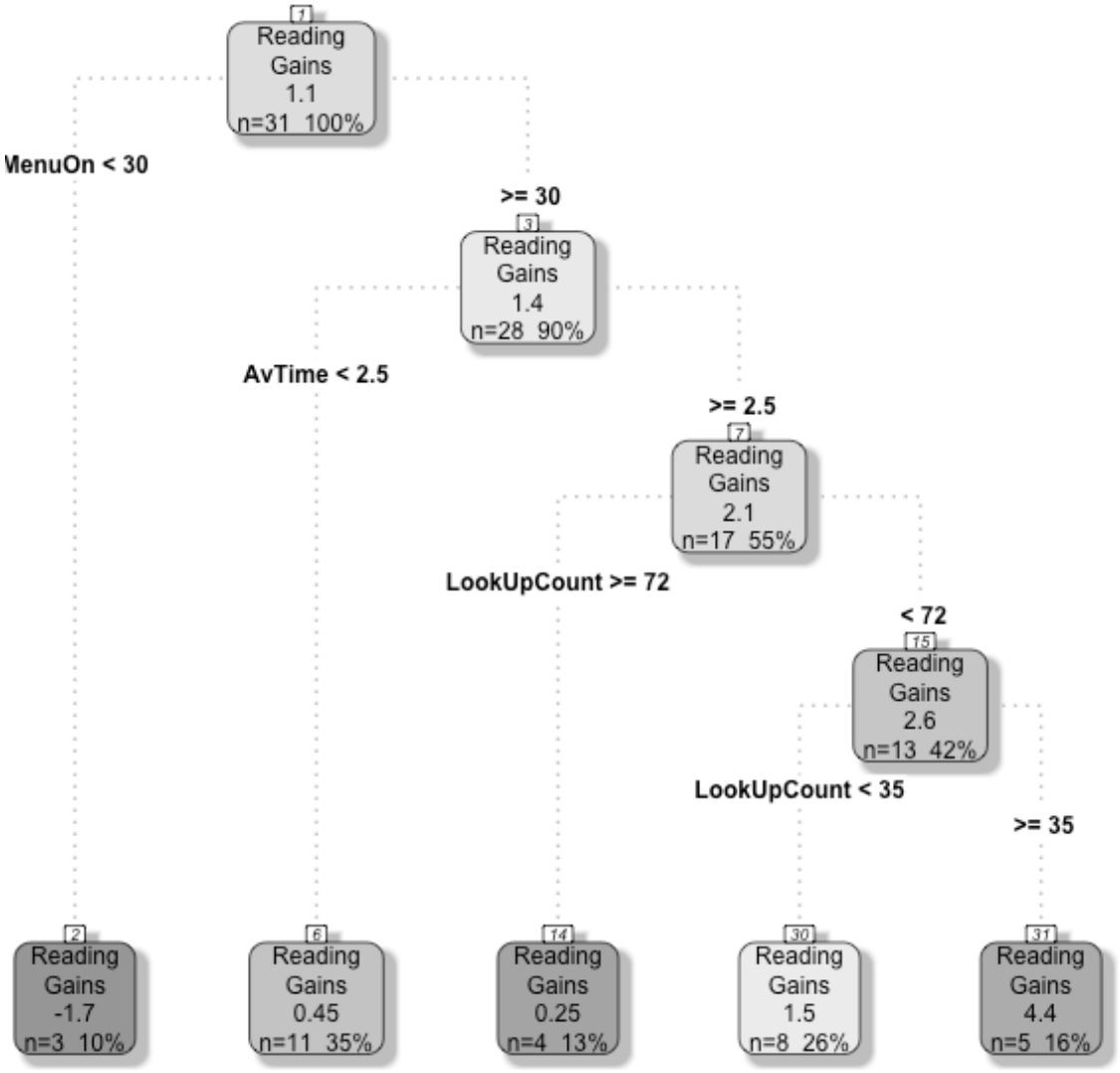


Figure 3.9

Reading Gains CART Analysis



Although this split is first, it should be interpreted conservatively given that only three students were in the less-than-30 node. The second split is on *AvTime.* This variable indicates the average amount of time spent reading a text. Students who spent on average more than 2.5 seconds reading a text (n = 17) reported on average a 2.1 gain on their

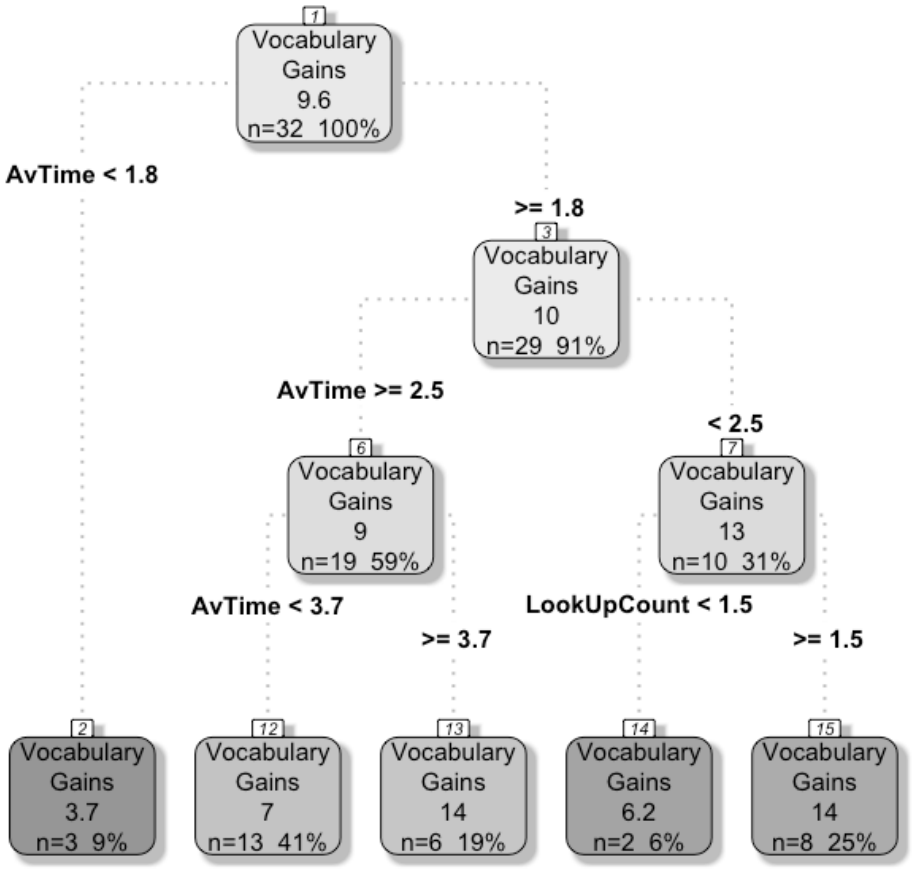
reading scores, compared to those who spent less than 2.5 seconds ($N = 11$) who reported only an average of .41 increase on their reading scores. *LookUpCount*, which is the number of times a student used the glossing tool, was used to split the last two nodes. Interestingly, those students spent more than 2.5 seconds on average reading a text and used the glossing tool less than 72 times, but more than 35 times ($N = 5$) reported an average gain of 4.4 points on the reading assessment. This suggests that students who used the menu options, spent longer time reading the text, and used the glossing function reported higher reading gains. Note that in order to build this decision tree, I first removed one student whose high reading gains were strongly influencing the decision tree. The analysis required a minimum of 2 splits, and was set to continue splitting each node as long as it produced an R change of at least 0.10. Given the small *N* in this study, this analysis was primarily concerned with identifying variables that were important for gains in Reading Comprehension.

The next CART analysis explores variables that are associated with vocabulary gains (see Figure 3.10). In this analysis, *AvTime*, or the average amount of time is used to determine most of the splits, with the *LookUpCount* variable deciding one split. While vocabulary gains are reported for groups in all final nodes, those who spent longer average times reading the text reported higher vocabulary gains. Similarly, those who used the glossing tool ($N = 8$) reported higher vocabulary gains on average ($M = 14$).

Because these vocabulary gains were primarily split depending on the variable *AvTime*, this variable was removed and another CART Analysis was created to explore other potential in-game indicators that were important for vocabulary learning. In this

Figure 3.10

Vocabulary Gains CART Analysis

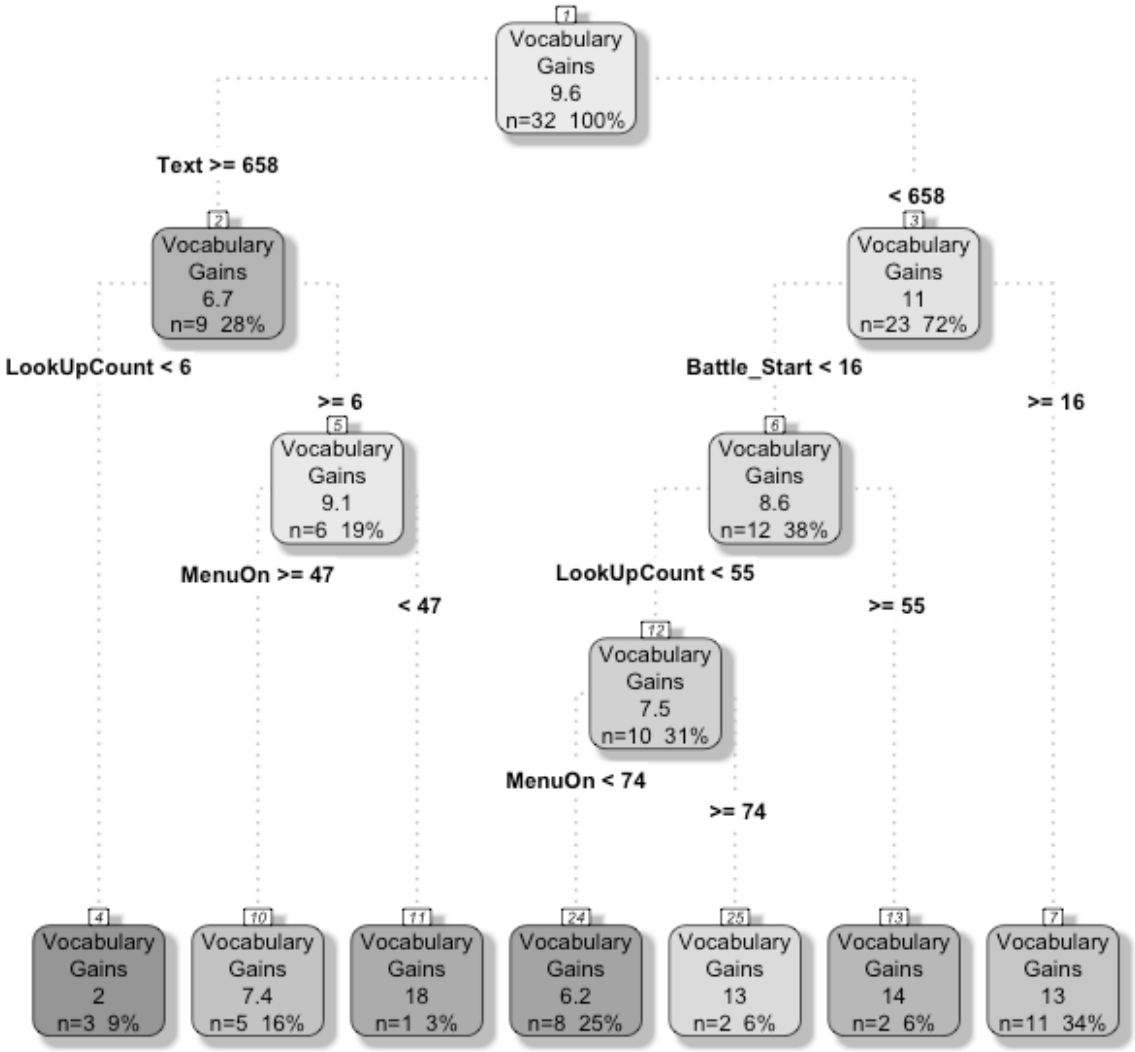


second CART Analysis for vocabulary gains, with average time being removed, the tree was first split based on *Texts* read (see Figure 3.11). Those who were exposed to less than 658 texts ($N = 23$) reported an average gain of 11 points on the vocabulary assessments. Those who were exposed to more than 658 texts ($N = 9$) only reported an average of 6.7-point increase on the vocabulary assessments. Those who read fewer texts but engaged in more than 16 battles ($N = 11$), reported an average gain of 13 points on the vocabulary assessments. Finally, in two separate splits with *LookUpCount*, those who used the

glossing tool had higher average gains on the vocabulary assessment than those who did not. In these two CART analyses, no cases were removed. The analyses required a minimum of 2 splits, and were set to continue splitting each node as long as it produced an R change of at least .06. For vocabulary gains, average time, the glossing function, the number of texts and battles appeared to be important.

Figure 3.11

Vocabulary Gains Decision Tree Without Average Time



Research question 2 asked, "*Are in-game playstyles associated with affective factors and/or vocabulary knowledge and reading comprehension gains, as measured by the pre-post assessments?"* To answer the second research question, I conducted a cluster analysis using four variables: *time, look up, texts*, and *number of battles*. These four variables were selected because they appeared in the decision trees which suggest their importance in relation to vocabulary knowledge and reading comprehension gains and because they are associated with gameplay styles. For instance, during the intervention it was noticed that some players carefully read every text, while other players tended to battle more, thus *battle_start* was deemed important as an indicator for gameplay style. Table 3.6 shows the four clusters that were identified as a result of the analysis. Before describing the unique aspects of each group in further detail, it is important to note that boys and girls were distributed evenly across all four groups, see Table 3.7. Also of note is the distribution of cluster by class, see Table 3.8. This is important because the *Battle Group* consisted almost completely of students from Class A, which may suggest that battling was a focus of Class A and also that playstyle was associated with the classroom environment and culture.

Table 3.6

Cluster Group Averages

Cluster group	*n*	Average time (seconds)	Average number of look up	Average number of texts	Average number of battles
Battle group	12	2.32 (-0.65)	21.91 (-0.43)	677.5 (0.82)	29 (0.98)
Not interested	9	2.37 (-0.59)	18.44(-0.54)	465.55(-0.54)	11.89(-0.72)
Close reading	10	3.94(1.40)	38.7 (0.04)	434.1 (-0.74)	14.2 (-0.49)
Vocabulary check	5	2.71(-0.16)	105.20 (1.94)	624.00 (0.47)	17.80 (-0.13)
Total	36	2.83	37.27	549.47	19.1

Note. Parentheses are standardized scores.

Table 3.7

Cluster Group by Gender

Cluster	Boys	Girls	Total
Battle group	5	7	12
Not interested	5	4	9
Close reading	5	5	10
Vocabulary check	2	3	5
Total	17	18	36

Table 3.8

Cluster Group by Class

Cluster	Class A	Class B	Total
Battle group	11	1	12
Not interested	3	6	9
Close reading	3	7	10
Vocabulary check	3	2	5
Total	20	15	36

The *Battle Group* contains 12 students and had the largest average number of texts as well as battles. This group of students also had the lowest average reading times and the lowest number of words looked up while reading. This group is given the name of *Battle Group* because these players likely preferred the battle aspect of the game.

Table 3.9 shows that this group had the highest average scores on the positive experience construct, suggesting that this group enjoyed the game more than others. They also had higher reading gains than two of the other three groups. The *Not Interested* cluster contains nine people. This group had the second lowest average read time and number of look ups, but the number of battles and texts read are also comparatively low

suggesting that this group did not engage with the text or the battles as much as the other groups. Table 3.9 shows that this group started with the highest anxiety scores on average, and the lowest vocabulary knowledge and reading comprehension scores, which may suggest that this group did not engage with the game as much because it was perceived as being too difficult. However, this group reported a relatively high score on the positive experience construct and vocabulary knowledge gains. Though it should be noted that with the lowest vocabulary knowledge average to start with, this group had more room to grow. Also, noteworthy is that this group had the lowest reading gains and also started with the lowest reading proficiency.

Table 3.9

Affect and Knowledge Score Averages by Groups

	Pre-vocabulary		Pre-reading comprehension		Pre-reading anxiety		Vocabulary gains		Reading gains		Positive experience	
Cluster groups	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Battle group	17.83	14.30	3.50	2.75	4.14	1.64	8.63	5.73	1.41	2.97	7.25	0.81
Close reading	16.94	11.15	3.40	1.92	4.30	0.92	11.33	5.37	1.44	2.01	6.35	1.67
Not interested	15.31	21.36	2.63	3.20	4.67	0.83	9.08	5.42	0.43	2.48	6.64	1.41
Vocab check	20.60	9.19	3.20	2.49	4.11	1.44	9.70	6.02	1.00	1.00	7.00	0.94

The *Close Reading* cluster contained 10 students. This group spent the highest average amount of time on reading, almost two standard deviations more than the class average. This group also had almost twice as many vocabulary word look ups on average than those in *Battle* and *Not-Interested* Groups. Given that this group spent more time reading the texts, it makes sense that this group read fewer texts, on average, and engaged

in fewer battles, on average, than the other three groups. Table 3.9 shows that this group had the lowest positive experience score on average, but their average was still over six out of eight. This group had the highest reading gains and vocabulary gains, on average, despite starting off with the highest reading comprehension scores, on average and the second highest vocabulary knowledge scores, on average.

Finally, the *Vocabulary Check* group had five students. This group looked up vocabulary words, on average, almost two standard deviations more than the entire group average. This group had the second highest average amount of time spent on text, and number of texts read. Given that this group not only spent a good portion of time reading the text and looking up words, but also was exposed to large amount of texts, this suggests that this group progressed far into the game by completing quests which require more reading. Further, given that this group had the highest pre-vocabulary knowledge average and the second highest pre-reading comprehension average, it would suggest that this group was full of high proficiency students.

After identifying the clusters, I then explored differences between the clusters on the outcome variables: positive experience, change in anxiety, vocabulary gains, and reading gains. I conducted four separate linear regressions to explore differences in learning and affect between the four groups identified in the cluster analysis. For the vocabulary, reading, and anxiety gains, prescores were entered into the model to control for ceiling effects. There was no prescore for positive experience so this was not entered into the model. No significant differences between these groups were found (see Table 3.10). The results of research question 2 illustrate how groups of students engaged with

the game in different ways. Much like one would expect, some students focused on the fun aspects of the game, while other students were more interested in the story or reading texts, and others were simply completing a task. The implications of these findings will be further explored in the discussion and conclusions section.

Table 3.10

Linear Regressions Comparing Outcomes by Cluster

Independent variables	Vocabulary gains regression coefficients (confidence intervals)	Reading gains regression coefficients (confidence intervals)	Anxiety change regression coefficients (confidence intervals)	Positive experience regression coefficients (confidence intervals)
Pre-vocabulary	0.214** (0.098, 0.329)			
Pre-reading comprehension		-0.522** (-0.801, -0.243)		
Pre-reading anxiety			-0.241 (-0.479, -0.002)	
Close reading cluster	2.898 (-1.154, 6.951)	0.173 (-1.569, 1.915)	0.272 (-0.466, 1.010)	-0.900 (-1.963, 0.163)
Not interested cluster	0.120 (-4.477, 4.717)	-0.598 (-2.580, 1.384)	-0.379 (-1.222, 0.463)	-0.607 (-1.788, 0.574)
Vocabulary check cluster	0.484 (-4.417, 5.385)	-0.573 (-2.676, 1.529)	-0.500 (-1.389, 0.390)	-0.250 (-1.571, 1.071)
Constant	4.814** (1.459, 8.170)	3.243*** (1.742, 4.743)	0.889 (-0.212, 1.990)	7.250*** (6.533, 7.967)
Observations	32	32	32	34
R^2	0.358	0.336	0.223	0.091
Adjusted R^2	0.262	0.238	0.108	-0.0001
Residual std. Error	4.688 (df = 27)	2.014 (*df* = 27)	0.853 (*df* = 27)	1.267 (*df* = 30)
F statistic	3.758* (*df* = 4; 27)	3.415* (*df* = 4; 27)	1.936 (*df* = 4; 27)	0.999 (*df* = 3; 30)

Note. Battle group cluster is reference group.

**$p < .01$

***$p < 0.001$.

Discussion

Past studies on games and L2 learning have explored how high and low proficiency students engage with a game (e.g., Rankin et al., 2006, how different game styles relate to learning (e.g., Collentine, 2011), and how gameplay is related to affect (e.g., Hwang et al., 2017). More research is needed in this area to not only explore how players engage with digital games in unique ways, but also how such data may inform classroom practices. In the present study, I explored gameplay patterns and types of actions that students engaged in (e.g., reading, battling) and how they were related to the learning and affect outcomes.

In the present study, CART analyses were used to identify variables that were important to both vocabulary learning and reading comprehension. In both analyses, the average amount of time that a student spent reading the text and the amount of times that the glossing tool was used were important variables in accounting for variance in learning gains. The average time that a student engages with a text makes sense as an important variable for a few reasons. First, it should be noted that interpreting the impact of time on text without considering the amount of texts being read is difficult. This is because some students may have been exposed to many texts because they were picking up a lot of items (e.g., these have a short, two-character text identifying the item) or they were skipping many texts. If the students took their time to read each of the texts, it's likely that their learning comprehension would increase, but these students are also likely to report being exposed to fewer texts. In addition, spending more time on the text allows for a better chance of vocabulary being processed and acquired. The glossing tools as

important indicators also makes sense as this is a tool that can be used voluntarily. Thus, when this tool is used, at some level, it can be deduced that the learner is purposefully lending more cognitive effort towards understanding the text. Finally, the effect of reading time on both vocabulary and reading gains is similar to the findings in Collentine's (2011) study in which reading time was significantly associated with syntactical complexity on a post-writing assignment.

The CART analysis of the reading scores also found that students engaging with options in the menu interface contributed to some variance in learning. These were moments when students accessed the menu to use or view an item and/or to explore the skills that were available to them. These likely led to some learning gains because each of these interfaces provided further linguistic support for in-game vocabulary and allowed for more exposure to text. For instance, in the menu, students can check a skill card that has a picture of the animal that can be summoned along with the pinyin for the animal. Further, these skill cards have icons for the items that the animals like and for the special skills that the animals possess. In terms of vocabulary learning, the amount of texts was also viewed as an important variable in a second, follow-up decision tree analysis. Interestingly, those who were exposed to fewer texts had higher, on average, vocabulary gains. Again, this is likely because those who were reading fewer texts, were also spending more time reading those texts and subsequently looking up words within the text.

To answer the second research question, I used a cluster analysis to group players into four groups that had unique playing styles. The first group was the *Battle Group*.

This group was unique because they had significantly more battles than other groups and more text exposure. This group also had the lowest average reading times suggesting that this group was largely interested in battling and picking up items. For many students in the class, battling was the best part of the game. As seen in Chapter 2, some students engaged in discussions with the researcher about how to defeat baddies. This group reported the highest average of game enjoyment and lowest average of prereading anxiety scores compared to other groups. This group also had the largest membership ($N = 12$).

The second group identified is the *Close Reading*. This group had the longest average reading times and the lowest average number of texts. This group appears to be the group that was more interested in the story. Unlike the *Battle Group*, the *Close Reading* group may not have been as interested in fighting. It is also possible that this group was more interested in using the game as a learning tool rather than as entertainment. This group reported the highest averages in both reading and vocabulary gains, though the lowest average on the positive experience construct. This group also reported the second highest vocabulary look up count.

The next group is the *Vocabulary Check* group. This group had the highest average of vocabulary word look ups, and was only comprised of 5 members. This group also had the highest pre-vocabulary scores on average. This is similar to earlier studies on glossing tools which found that students with higher proficiency levels used the tool more (H. Lee et al., 2019; Poole & Sung, 2016). This could also be explained by Hsiao et al. (2017) in which they explored patterns in how learners explored a virtual world. They noted that students with higher L2 proficiencies were more purposeful in the types of

vocabulary words that they clicked on to learn.

The last group was deemed the *Not Interested* group. The three aforementioned groups all had an area in which they excelled or set them apart from the other groups. For example, the three other groups may have enjoyed the game because they got to battle, they found the story interesting, or they learned vocabulary. The *Not Interested* group had the second shortest reading time average suggesting that they did not engage in the reading. They also reported the second lowest look up average and text exposure average. This again suggests that they did not engage in reading the text as much as their other classmates. While they did engage in more battles than groups three and four, they battled less than half as much as the battle group. It could be that these students had low reading comprehension and therefore had trouble accessing the content or meaning in the game. After further analysis of this group, three of the students were those who missed at least two days of the study, two other students were identified as having technical difficulties at the beginning of the project. These are not trivial items and are real issues that educators have to deal with when integrating games into the classroom. It suggests that technical issues can affect learning and how students engage with the game.

Finally, differences between these group in terms of learning and affect were not significantly different. This may have been due to a lack of power, but if not, this is a positive finding in that it suggests that players, regardless of playstyle, learned vocabulary and improved their reading as a result of the intervention. In other words, the game *Legend of the Dragon* allowed students to play and learn in their own way. This highlights an advantage of open-ended games. Unlike other games with controlled, linear

progression from level-to-level, *Legend of the Dragon* has multiple entry points for students to engage with L2 content in ways that are meaningful to them.

Conclusion

Research using educational data mining approaches in L2 gaming contexts is rare. As noted by other researchers, applying such approaches to explore L2 learning in digital environments may be beneficial to educators (Godwin-Jones, 2017; Reinders, 2018). Further, collecting and analyzing data from digital games using EDM approaches may allow educators to visualize and capture learning as it happens in context, much like Yim et al.'s, (2017) study which used visualizations to explore the collaborative writing process. Further, research on EDM and L2 digital games has primarily focused on students at the university level (e.g., Foung, 2019; H. Lee et al., 2019), with no other studies exploring how L2 elementary learners engage with a digital game in the classroom. Research using EDM techniques at the elementary levels is important as it may inform teacher practices involving digital games for L2 learning.

There are several ways that information from the present study could be used by teachers to inform their classroom instruction when using the game in the classroom. Creating a teacher dashboard that provides visualizations of the data is one way to use game data. Some examples include providing visualizations of the vocabulary words accessed (or not accessed), during game play that teachers could use to guide either activities supplemental to game play or post game play. Information on clusters could be used to find information about how students are accessing the game and what in game activities may lead to learning gains. In addition, by identifying learner gameplay styles,

an instructor may decide to organize group activities around the game by leveraging the groups identified in the cluster analysis. For example, in the present study, the instructor may want to pair some of the *Not Interested* students with more proficient readers to provide additional support while playing the game. Further, these findings may be beneficial to future game designers by providing insight into the role of time spent reading a text, use of the menu, text exposure and glossing tools. Specifically, this study found that reading gains on average were higher for those who spent more time reading the text, and using the glossing tool. Finally, before such data and analytics are provided to educators, future research should explore teacher perceptions and beliefs about data and visualizations derived from in-class activities.

Limitations

The present study has a few limitations and thus the findings reported should be interpreted conservatively. First, although gameplay provides several data points which is ideal for data mining approaches, the total number of participants in this study is small. To further confirm some of the trends associated with learning in this study, research with larger populations is needed. Second, there was no relative qualitative data associated with the analysis in this study. While I did collect affect data via surveys, video data or interview data that supports the conjectures made about the groups identified in the cluster analysis would have been ideal.

CHAPTER 4

DEVELOPING AND VALIDATING STEALTH ASSESSMENTS FOR AN EDUCATIONAL GAME TO ASSESS YOUNG DUAL LANGUAGE IMMERSION LEARNERS' READING COMPREHENSION

Abstract

The present study draws from the evidence-centered design framework to build and evaluate stealth assessments within a digital game to assess young second language learners' reading comprehension. Assessing reading comprehension is a difficult and complex task. Using digital games as a means to collect data to assess learners while they engage with texts may be a viable solution. Further, by assessing learners in a low-risk environment that also promotes learning, stealth assessments may become a better alternative to traditional assessments. The present study uses Bayesian belief networks to leverage time on task, use of a glossing tool, and student vocabulary knowledge as a means to assess readers as they engage with in-game texts. This study found that scores produced by the Bayesian belief network is significantly correlated with two external measures of reading comprehension.

Introduction

Foreign/Second Language (L2) educators have long recognized the value of integrating games into instruction (Baltra, 1990; Cornillie et al., 2012; Hubbard, 1991; Prensky, 2001). Over the years, researchers have argued and provided evidence that

digital games afford several educational benefits for L2 learners including a more enjoyable learning process (Becker, 2007; Gee, 2003; Prensky, 2001), motivation to persist in learning the L2 (Hayes, 2005; Prensky, 2001; Warschauer & Healey, 1998), a highly contextualized and interactive learning environment (Gee, 2003; Morton et al., 2012; Presnky, 2001; Sørensen & Meyer, 2007; Vogel et al., 2006), opportunities for collaboration and meaningful interactions (Dalton & Devitt, 2016; Peterson, 2011; Warschauer & Healey, 1998), and immediate feedback in context (Cornillie et al., 2012). Furthermore, these benefits have been shown to promote vocabulary learning (e.g., Ansteeg, 2015; Bytheway, 2014; Yudintseva, 2015), a willingness to communicate (e.g., Reinders & Wattana, 2014), writing skills (e.g., Coleman, 2002; Palaiogiannis, 2014; Suh et al., 2010), among other L2 skills, while also reducing anxiety associated with learning an L2 (e.g., Hwang et al., 2017).

Given these benefits, one of the often overlooked and under researched advantages of utilizing digital games for L2 learning is the potential to assess learners while they engage with and use the L2 in highly contextualized environments, without the pretext of an assessment (Mavridis & Tsiatsos, 2017). Traditional assessments (e.g., fill in the blank, multiple choice items) have been criticized both due to their lack of contextualization (Baker et al., 2016; Shute & Wang, 2015) and because they separate assessment from learning via the pre-post assessment paradigm (Clarke-Midura & Dede, 2010; Halverson & Owen, 2014). Digital games can be leveraged to track player actions, in-game choices, and behavioral patterns, which may act as indicators for L2 proficiency and/or growth. Perhaps more importantly, data collected in digital games provide

multiple observations of student performance which is needed to make an adequate assessment of student learning and/or competencies (Clarke-Midura & Dede, 2010). Furthermore, such measures can be collected while–not after– learners engage in tasks that are meaningful to the learner. Finally, such data may also prove to be useful for in-classroom educators to inform their practices.

In other educational subject-content areas (mathematics, biology, physics, etc.), researchers have utilized log data files from digital games to collect evidence linking in-game behavioral actions with cognitive choices (e.g., Gibson & Clarke-Midura, 2015; Gobert et al., 2012), changes in a learner's conceptual understanding (e.g., Martin et al., 2015), and/or affect change (e.g., Halverson & Owen, 2014). While these approaches are considerably new in educational research, they have been almost non-existent in studies investigating L2 proficiency and development. L2 studies using in-game log data files tend to use such data to track student responses on multiple choice-like questions (e.g., Erhel & Jamet, 2016) or to collect and later analyze L2 text produced during in-game chats (e.g., Collentine, 2011). However, log data files could be used to track players' interactions with new vocabulary, use of in-game support while reading an L2 text, and other strategies employed while attempting to overcome linguistic challenges. Subsequently, such data may not only facilitate the creation of better models for L2 learning, but they may also allow better predictions or assessment of language proficiency. However, before this can happen, researchers must first identify in-game behaviors that are both trackable with log data and representative of L2 learning and/or proficiency. Identifying in-game behaviors includes both identifying gameplay

tendencies (e.g., amount of time spent reading in-game texts, types of in-game activities engaged in) and designing and evaluating activities that elicit behaviors representative of L2 learning or proficiency. Finally, these assessments must then be validated before they are used in meaningful ways.

In the present study, I applied game-based assessment approaches being implemented in other educational settings to an L2 learning context. To do this, I designed and built a digital game for students in a Chinese dual language immersion (DLI) classroom to promote vocabulary learning and reading development (see Chapter 2). Within this game, student data related to the use of vocabulary support, reading in-game texts, and gameplay behaviors were collected. This data was used to construct and validate a Bayesian Belief Network to automatically assess young Chinese DLI learners' reading comprehension proficiency. In the sections that follow, I first provide a review of the literature, discussing the L2 reading process and research approaches to assess L2 reading comprehension. Then I review literature around game and virtual assessments and make an argument for game-based assessments. Next, I introduce the evidenced-centered design framework that I used to guide the design of the stealth assessments. Then I introduce the methods, including a brief description of the game used in this study and a detailed description of how the stealth assessments and Bayesian belief networks were constructed. Finally, I discuss the implications of the stealth assessments designed in this study for L2 contexts.

Literature Review

Reading in the L2

Reading is the process of extracting information from written text. This means readers must decode orthographic, phonological, semantic, morphological, and cultural knowledge from text. Researchers have argued that readers do this by mapping knowledge acquired orally from a language to the associated writing system (e.g., Perfetti, 2007). Although early research relied heavily on studies involving L1 readers to inform hypotheses and conjectures about L2 reading processes (Bernhardt, 2005; Koda, 2007), L2 reading differs substantially from L1 reading. One of the most obvious ways that L2 reading differs from the L1 is that the reader has two language systems. In other words, when one engages in L2 reading, the L1 is often used as a resource for understanding the text. Depending on linguistic, syntactic, and cultural proximity of the L1 to the L2, the impact of a second language can vary significantly on how one learns to read in the L2 (Koda, 2007). In Bernhardt's (2011) compensatory model of L2 reading, she argues that reading comprehension ability is comprised of a reader's L1 literacy skills, L2 language knowledge, some other unexplained variance which may include background knowledge, life experiences, and metacognitive strategies.

Another difference between L1 and L2 reading is that L2 learners, especially novice and intermediate learners, do not have the advanced oral language skills of a native speaker to rely on when learning to read (Koda, 2005; Yorio, 1971). Similarly, there is typically a large gap in the amount of vocabulary knowledge that learners have when reading texts (Clarke, 1980; Grabe, 2014). These two differences result in more

bottom-up processing by L2 readers, and more specifically novice and intermediate L2 readers, when reading a text in their target language (Clarke, 1980; Grabe, 2009, 2014; Koda, 2005). Bottom-up processing represents the decoding of individual words first, and then after processing each word, trying to make sense of the statement (Hinkel, 2017). In contrast, top-down processing involves the use of prior knowledge and understanding of the text to make inferences and predictions about the text (Tsui & Fullilove, 1998). Top-down processing is done when less cognitive effort is used for decoding words. In other words, once learners can automate the process of recognizing words and their meanings in a sentence, they can begin to consider the meaning of the text on a global level.

Similarly, L2 reading has been described as containing lower- and higher-level processing where lower-level processing refers to reading the text at a superficial level (Alderson et al., 2015). In contrast, higher-level processing refers to many of the meta-cognitive activities that take place while reading, such as making inferences about the author's intent, responding to the arguments or statements made in text, and so on (Koda, 1992). These two perspectives on reading are similar, but what is important is that many novice L2 readers spend much of their time using bottom-up strategies for lower-level processing (Clarke, 1980; Grabe, 2009, 2014; Koda, 2005). This is important for the present study because to assess L2 learners' reading skills, one must understand what type of reading they are engaging in. While researchers generally agree that L2 readers will engage in a mixture of processes at some level (Alderson et. al., 2015), given that novice learners are so focused on decoding vocabulary, their high-level processing is typically limited.

Assessing L2 Reading

When thinking about L2 assessments, it is first important to determine what is being assessed (Roever, 2001). In other words, is the assessment aimed at measuring L2 speaking or reading, or is the goal to assess performance on a communicative task. Next, one should consider the purpose of the assessment. Shrum and Glisan (2010) state that assessments used in L2 classroom settings are used to understand the learning process, identify learner struggles, and track student progress. They further argue that this information can then be utilized to improve L2 instruction. Assessments have also been used to make claims about students' achievement or proficiency (e.g., Hadley, 2001). An assessment focused on achievement is one which assesses a learner's knowledge of a targeted skill or knowledge, while proficiency assessments assess a learner's overall L2 ability (Hadley, 2001). The current study aims to assess learner reading proficiency using their engagement with in-game texts.

Another distinction to make between assessments is whether or not the assessment is authentic. An authentic assessment has been defined as one that resembles the tasks and activities that one engages with in the real world (Wiggins, 1998). Traditional assessments are contrary to authentic assessments in that they divorce the L2 concept from their real-world use. This is often done with achievement tests when educators attempt to isolate a specific skill or language to be tested. Using authentic assessments is argued to be beneficial to learners because measuring learners based on real-world use of a language not only provides a better evaluation of what a learner can do with a language (rather than what they know about a language), but it also creates a

washback effect, in which teachers adjust their pedagogy to reflect the test (Lantolf & Poehner, 2003). In other words, if assessments are divorced from real-world use of the L2 then it is likely that L2 teachers will likewise change their pedagogical approaches to mimic the assessment. Shrum and Glisan (2010) provide a list of guidelines for creating authentic L2 assessments that include:

- Contextualizing the assessment
- Meaningful communication
- Elicit performance of some type
- Encourage divergent responses and creativity
- Adaptable

In terms of L2 reading comprehension, research has also shown that task type greatly influences student outcomes (J. Lee, 1987; Wolf, 1993). Specifically, multiple choice questions, fill-in-the-blank, essay response items among others all may elicit different skills (J. Lee, 1987; Wolf, 1993). J. Lee and VanPatten (2003) further illustrate the complexity of measuring reading comprehension when they say, "all attempts to test and evaluate comprehension are problematic because the process is internal to the reader" (p. 262). In other words, it's impossible to determine if the student is actually reading or comprehending what is being read because it is happening cognitively. Shrum and Glisan (2010) argue that when assessing reading comprehension, items should reflect what students do in the classroom and further that "reading should be constructed to encourage learners to read more" (p. 262). In the current study, the game-based assessments are seen as authentic assessments as they assess students as they engage in the real-world activity of playing a game.

Past research on assessing L2 reading skills has followed one of three generic

approaches. One of the first approaches is an assessment that focuses on the product, or what students understand from a text (MacMillan, 2016). Researchers have then tried to make inferences about certain skills and/or characteristic that are associated with high and low performers. This approach has been criticized largely because it has been pointed out that what learners take from a reading can vary significantly, and there is no clear method for distinguishing value of one product over another (Alderson, 2000). After a focus on reading products as means to assess reading comprehension, researchers shifted towards a process approach in which the interaction between learner and text is examined. More specifically, the learner is assumed to come to the text with knowledge and experiences that will affect how a text is read, what content receives attention, and the perceived difficulty of the text. This approach has led to much valuable research (e.g., Block, 1992; Kern, 1994; Leow & Morgan-Short, 2004); however, it has been criticized because the approaches to extract evidence of reader interaction with the text is also divorced from the natural reading process (e.g., read aloud protocols; Stratman & Hamp-Lyons, 1994; Yoshida, 2008). More recently, much research has focused on factors that affect reading difficulty and discrimination using psychometric measures (MacMillan, 2016). By determining item difficulty and discrimination features within a reading comprehension prompt, much can be inferred about student reading abilities depending on how the student performs on a reading comprehension item. The present study takes the position that many of the challenges currently involved in assessing L2 reading can be resolved via game-based assessments.

Assessments in Games and Virtual Environments

Although research investigating the design and validity of assessments in games and virtual environments is still in its infancy, several studies have measured learning and development using game and simulation data in a variety of ways. Some studies have used time and frequency variables as predictors for learning (Chin et al., 2016; DiCerbo, 2014; Galaup et al., 2015; Liu et al., 2016; Shute & Wang, 2015), while other studies have examined the significance of sequence of choice and user path selections in virtual environments (Halverson & Owen, 2014; Loh & Sheng, 2015). In addition, some researchers have made inferences about learning outcomes based on player performance on tasks designed to elicit specific skills related to the learning outcomes (Baker & Clarke-Midura, 2013; Gobert et al., 2012; Quellmalz et al., 2013).

Much of the research using frequency and time variables tends to count the number of times a player engages in an action and/or the amount of time spent on the action. These frequency- and time-based variables are then used as predictors for learning gains on external assessments. In a study investigating the effect of a serious game on data literacy and visualization skills, Chin et al. (2016) found that frequency and duration of in-game tasks significantly predicted learner gains on external assessments. Similarly, Galaup et al. (2015) found that in-game successes (i.e., in-game gold collected) and time spent playing the game outside of class while playing a serious game designed for mechanical engineering students significantly correlated with final exam scores. In another study, Liu et al. (2016) used an Alien encounter-themed game to teach problem solving skills. However, Liu et al. were interested in comparing how high and low

performance students used the resources in the game. They tracked frequency and duration of tool use by players and found that high performers were better at selecting and using the proper tool for the task. Using vernacular games, Shute and Wang (2015) tracked several in-game actions of students playing either Portal 2 or Lumosity, including number of levels completed, average time spent on a level, and several measures of attention provided by Lumosity. They found that indicators in Portal 2 significantly predicted post-test outcomes while those from Lumosity did not. Finally, DiCerbo (2014) used in-game analytics including, time spent on task and number of tasks completed, as indicators for persistence in an activity. In L2 studies using digital games, some researchers have also investigated the relationship between time and frequency variables on learning outcomes. Collentine (2011) found that the amount of time participants spent on either reading a text or a reading question was significantly associated with outcomes on a writing assessment after playing a digital game designed for second language learners of Spanish. In another study investigating the use of Everquest II as an English language learning tool, Rankin et al. (2006) found that more advanced speakers generated more chat messages than intermediate and novice speakers.

Research examining player path and sequence of actions have examined the relationship between combinations of in-game behaviors and outcome variables. For instance, Halverson and Owen (2014) used log data to identify players who seemingly randomly clicked buttons in a game designed to teach stem-cell science and found that this behavior was associated with poor learning outcomes on an external pre-/post-assessment measuring student knowledge about stem-cells. However, they also found that

learners who made it to the "boss level," or end of a level, experienced significantly higher gains on the post-test. In a study that tracked player paths in a military game, Loh and Sheng (2015) developed statistical measures to compare expert paths to player paths. Their study illustrates how player sequence through tasks can be used to identify proximity of expert among players. In a L2 study using a digital game, Hwang et al. (2017) compared in-game behavior sequences by EFL learners with high and low anxiety (as determined by a pretest). They found that learners with higher levels of anxiety engaged in more 'complex' forms of learning within the game than less anxious students. In other words, students with less anxiety used in-game supports and then completed a task, while students with higher anxiety engaged in a more iterative process that included checking in-game supports multiple times and failing a task before completing the tasks. They concluded that high anxiety learners reviewed more vocabulary cards and acquired more relevant knowledge prior to completion of tasks. Finally, in studies that attempt to match in-game behaviors to educational competencies, researchers first identify indicators within the game and then typically use Bayesian statistics to establish a link between the indicators and the targeted skill or competency. For example, Gobert et al. (2012) identified several skills related to scientific inquiry, and then illustrated how tasks in their virtual experimentation lab elicited the targeted skills. Last, they provide a framework for assessing and automatically scoring each of the skills with a binary code. In another study, Quellmalz et al. (2013) compared validity measures of virtual assessments in three different learning environments (i.e., static, active, and interactive). They found that interactive environments provide learners with better estimates of

assessments in regard to scientific inquiry. Baker and Clarke-Midura (2013) used Bayesian Knowledge Tracing to explore how in-game indicators predicted both correct responses on a task designed to elicit scientific inquiry, and the causal explanations developed by the learner for the response. They argue that such assessments can be used to assess a learner's zone of proximal development. In terms of L2 studies, Cornillie et al. (2017) designed tasks within a virtual world to allow EFL participants to practice applying their knowledge of English pragmatics to a meaningful and contextualized environment. The authors were interested in exploring the types of feedback that learners preferred while playing the game. While the authors concluded that more explicit feedback was more effective and preferred, they did not validate their tasks as assessments. In another study exploring vocabulary learning in a simulation software, Hsiao et al. (2017) collected player log data to analyze patterns in how students interacted with new vocabulary words. It should first be noted that the students were given a specific task of learning 30 words and there were no clear gaming elements in the software. Through data visualization techniques, the authors found differences in learning strategies between high- and low-achieving learners. Specifically, they found that low-achieving learners tended to click on vocabulary words randomly or use the nearest neighbor approach in which they simply clicked on vocabulary words that were close to each other. In contrast, the high-achievers tended to use a strategy to cluster similar vocabulary words together to facilitate learning. This study is particularly important for the present study for a few reasons. First, it is one of the few studies that has used data mining techniques to explore L2 player interactions via digital software when learning

Chinese. Second, this study confirms that behavior in the software was dependent on L2 proficiency. Finally, given that participants were told to learn 30 words, and there were no game elements embedded into the software, there is still a need to explore how learners engage with L2 text in game when they are more focused on playing the game rather than learning vocabulary.

Again, research on game and virtual assessments is still a nascent field. However, research involving game and virtual assessments in an L2 context is almost non-existent, especially for non-English languages.

The Present Study

As presented in the literature review, assessing L2 reading comprehension is a complex process because there are many variables that are associated with reading and the reading process is internal to the student. Assessing L2 reading within a digital game at first glance appears to be more complex, especially in a game that allows players autonomy in how they play (see Chapter 3). In the digital game used in the present study, although several dialogues prompt learners with a question to respond to, the answers are not necessarily representative of failed comprehension. For instance, if a noncharacter player (NPC) asks a student for help, the student may choose to not help the NPC. This does not necessarily indicate failed comprehension. Similarly, for other dialogues that prompt a response, students may simply be interested in exploring how atypical responses affect the system. Further, given the unique ways that learners play a game (see Chapter 3), the type and amount of text that a player is exposed to may vary significantly.

While this does appear to be more complex, reading within a digital game allows data related to student engagement with a text to be collected. In the present study, the amount of time spent on the text, the use of a glossing tool within the text, and a student's knowledge of in-text vocabulary is leveraged to make inferences about whether or not a text was read and understood. To leverage these data points, I used a stealth assessment approach to embed assessments within the digital game play.

Assessment Design Framework

The present study employed a Stealth Assessment approach (e.g., Shute & Kim, 2014), which is based on the Evidence Centered Design Framework (ECD). Stealth Assessments are evidence-based approaches that unobtrusively assess students while they are interacting with digital games or virtual environments (Shute & Ventura, 2013). This approach was used for three reasons. First, it provides a rigorous model for developing an assessment in terms of evidentiary arguments (Mislevey et al., 2003; Mislevy & Haertel, 2006). Second, ECD was developed for complex assessments such as this and allows for the modeling of game data as evidence for students. Further, this framework ensures that evidence collected and analyzed is aligned with the goal of the assessment.

Evidence Centered Design

Evidence Centered Design (ECD) is a multilayer approach comprised of five layers: (1) domain analysis, (2) domain modelling, (3), the conceptual assessment framework (CAF), (4) assessment implementation, and (5) assessment delivery. In layers 1 and 2, the focus is on the purposes of the assessment, the nature of knowing, and

structures for observing and organizing knowledge. In the third layer, assessment designers focus on the competency model (e.g., what skills are being assessed), the evidence model (e.g., how are skills measured), and the task model (e.g., situations that elicit the behaviours/evidence). These aspects of the design are interrelated. The models promote a shift in assessment design towards a focus on the process of task design and how such tasks relate to statistical interpretation, rather than simply analyzing and interpreting data out of context (Mislevy et al., 2003).

Stealth Assessments

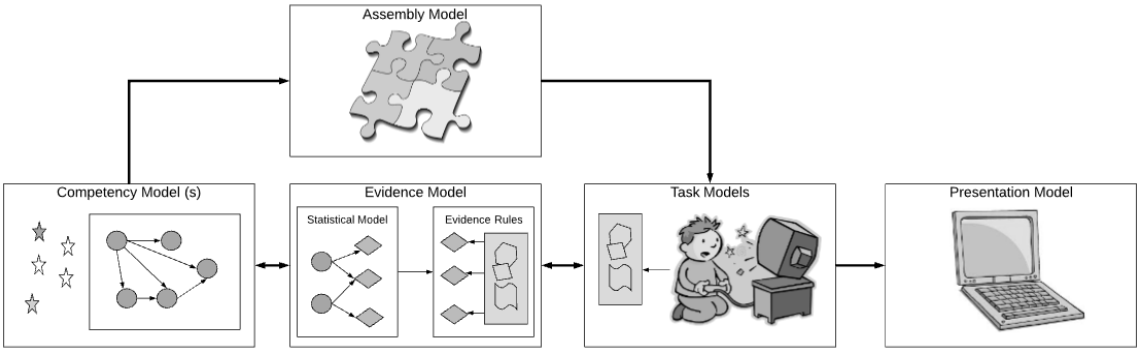
Stealth Assessments are designed using ECD's framework. Of particular interest is the third layer, the Conceptual Assessment Framework (CAF). The CAF contains several models that work together to answer the questions "what attributes are to be measured?" and "how do we store them?" (Shute & Kim, 2014, p. 24). This layer focuses on the assembly of the entire assessment and the processes needed to implement it by generating a test blueprint that specifies the scoring system, statistical models, and delivery processes (Mislevy & Haertel, 2006; Mislevy & Risconcente, 2006). Figure 4.1 provides a visualization of the five components that make up the CAF and how they are interconnected.

What Are We Measuring?

The first model is the competency model, which has also been called the student model. This model answers the question of what the assessment is measuring and represents the claims that are being made about students' proficiencies (Mislevy &

Figure 4.1

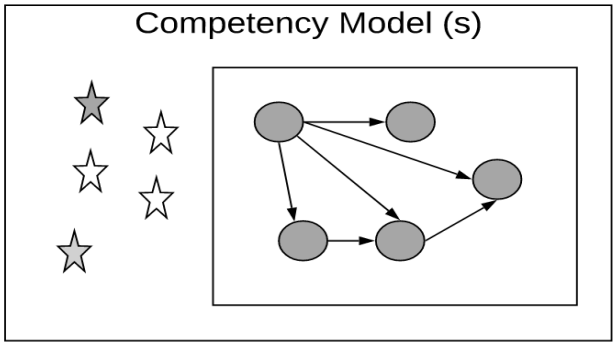
Conceptual Assessment Framework



Risconcente, 2006). In a game-based assessment, as in the present study, the competency model is constantly updated based on how students interact with the in-game tasks, see Figure 4.2 (based on the evidence model; Shute & Ventura, 2013).

Figure 4.2

Competency Model



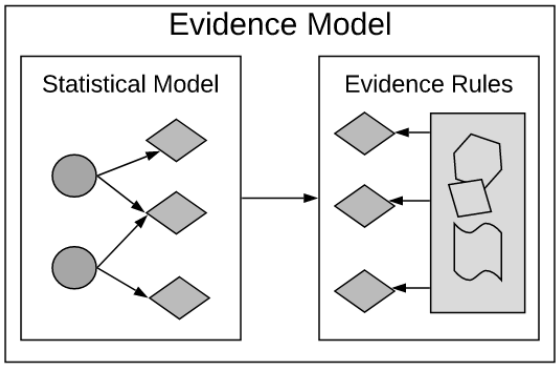
How Do We Measure It?

The second model is the evidence model. The Evidence Model provides the technical details on how the assessment measures the Competency model. The evidence

model contains two parts, the evidence rules and the statistical model. The evidence model refers to how information is decoded, in this case from a computer, to match with the student variables, see Figure 4.3. For instance, in the present study it would refer to how time on reading tasks was calculated and eventually discretized for analysis. The statistical model explains how evidence is associated with the competencies outlined in the competency model. Again, for a simple example, the statistical model might state that 80% or higher on a grammatical quiz indicates mastery of the grammatical concept. However, for this study and for many other studies that employ the ECD framework (e.g., Kim et al., 2016; Shute et al., 2017), Bayesian belief networks are used to describe the probability of reading comprehension given the evidence of student performance on tasks with varying degrees of knowledge and difficulty.

Figure 4.3

Evidence Model



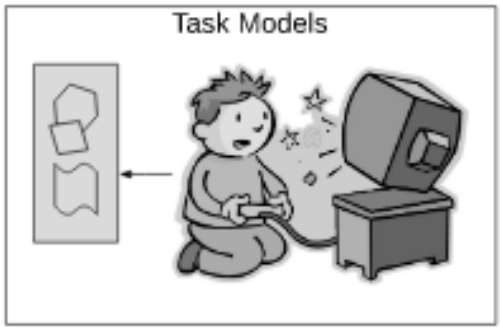
In What Situation(s) do We Measure it?

The third model is the task model, see Figure 4.4. This model focuses on the design of tasks that will elicit the competency variables above. Simply stated, this model

describes where and how we measure the targeted skills and/or knowledge. Tasks should be described by how they are presented, types of outcomes elicited, and in terms of variables associated with the task features (e.g., difficulty). Again, in a simple example, this model might describe how an open-ended essay question is better at eliciting grammatical knowledge when compared to a multiple-choice question. Specifically, this model describes the prompt given to the learner, then the types of responses that may be elicited from this prompt, and finally how features (e.g., open-ended vs. multiple choice) will impact the results produced. Subsequently, this model considers all possible behaviors that may be elicited from a task and if those behaviors can then be mapped on to the targeted competencies. In the present study, the tasks are short reading prompts that appear throughout the game. There are some reading prompts that require a response and some that do not. All reading prompts provide users with the option to use a support system that provides the Pinyin for targeted words. The primary behavior that will be elicited is time spent on reading and whether or not the support system is used. However, other variables that define the task features will also be collected including length of text,

Figure 4.4

Task Models



whether or not the learner has vocabulary knowledge associated with the text, the amount of times the text has been seen, and whether or not the text prompts a response.

How Much Do We Need to Measure It?

The assembly model describes how the student, evidence, and task models work together to provide sufficient information for a valid assessment (Almond et al., 2014). Most importantly, the model specifies how much information is needed from each variable in the competency model in order to represent the breadth and diversity of what is being assessed (Mislevy & Risconcente, 2006). This is particularly important for GBA because of the variability in how students play a game. For instance, some text may be seen by only one player, while others are seen by all players, multiple times. The last model is the presentation model. This model describes how the game is presented and played on certain devices (e.g., laptop, tablet) and how it is presented to the learners (e.g., as an assessment or as a game to play).

The primary focus of the present study is to develop a stealth assessment, a specialized implementation of the ECD framework that involves embedding assessments into a learning environment (Shute, Ke, & Wang, 2017). The stealth assessment will be used to assess young Chinese DLI learners' reading comprehension. The present study was guided by the following research question.

1. *Do students' performances on in-game task, as determined by the stealth assessments, correlate significantly with external, paper-based, post-reading measures?*

Method

Participants and Setting

Data were collected in two sixth-grade elementary Chinese DLI classroom settings. The two classrooms had 19 and 21 students respectively. Stratified randomization with gender as a factor was used to assign learners to each classroom at the beginning of the school year. Students ranged in age between 10 and 12 years old. Four students (1 in the Class A, 3 in Class B) did not sign the consent form. In addition, two students in Class B did not take the pre-test, and one student in both Class A and B did not take the post test. Thus, although I had consent forms for 36 students, there was only complete data from 32 students, yet for some of the analysis used in this dissertation all 36 of these students' data were utilized.

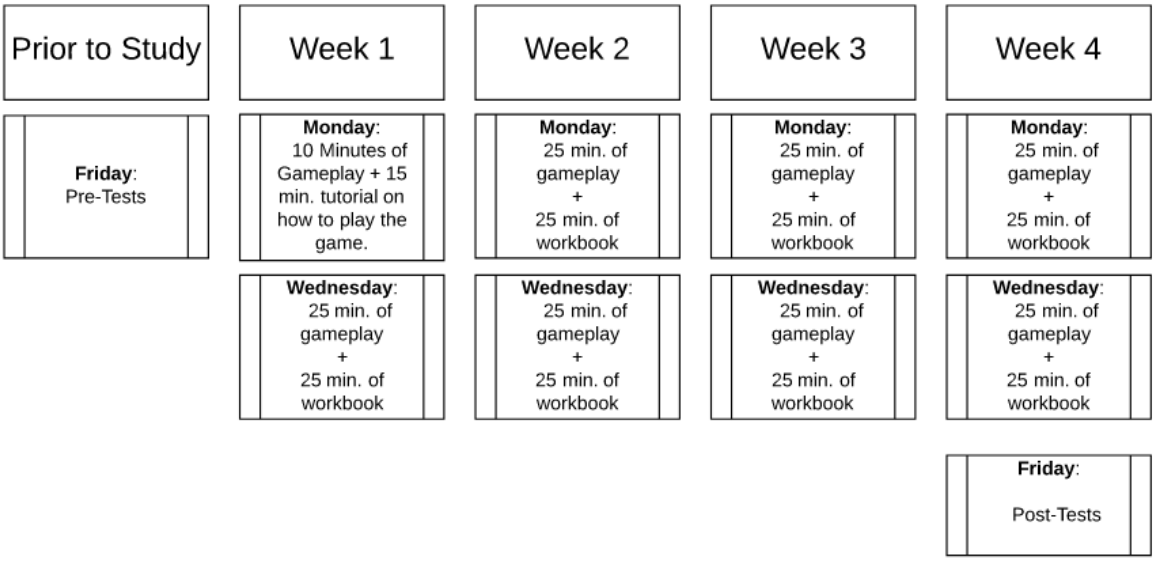
Procedures

The game was played on a set of MacBooks that were provided by the research team. All data was uploaded to a Mongo Database after each gameplay. Eleven computers were used which meant that only half of the class could play the game at a time. Thus, while one half of the class played the game, the other half completed a workbook that reviewed in-game vocabulary and phrases (see Appendix B). This workbook largely consisted of character writing tasks, vocabulary matching exercises, and oral discussion activities about the game. This workbook was designed to scaffold learner interaction with the game by providing linguistic support via character recognition tasks and gameplay strategy via oral discussions. Half of the class completed

the workbook while the other half of the class played the game. The half of the class that completed the workbook was first given a brief introduction to the exercises that they needed to complete, and then they completed the exercises on their own. The students played the game for four weeks, twice per week, and approximately 25 minutes per session (see Figure 4.5).

Figure 4.5

Research Schedule



Data Collection

Data for this study came from four sources: pre- and post-vocabulary and reading comprehension assessments, teacher evaluations of student reading comprehension, and log files that captured in-game actions, texts read, and choices made.

Vocabulary Assessment

The pre- and post-vocabulary test consisted of 45 words that could be found in the

game. All of the vocabulary words selected were words that could be checked via the pinyin support system. Although it should be noted that this is not a comprehensive list of the vocabulary words that could checked. On the vocabulary assessments, learners are presented with the character and then learners were first asked to declare if (a) they know the word, (b) they think they know the word, or (c) they are just guessing. Next, learners were prompted to enter the pinyin and the English translation. The vocabulary assessments can be seen in Appendix B. These assessments were scored by awarding 1 point for correct pinyin, and 1 point for correct English definitions. Further, half points were awarded for partial answers. For example, if a learner correctly identified the correct pinyin or English for one of the characters but not both half points were awarded. Awarding partial points for vocabulary knowledge was viewed as valuable given past research that has noted the non-linear trajectory, and the partial accumulation of vocabulary knowledge that occurs through incidental learning while reading L2 texts.

Reading Comprehension Assessment

The external reading comprehension test was adapted from the Chinese YCT test (http://english.hanban.org/node_8001.htm), which is an official Chinese proficiency assessment developed by the Confucius Institute and used regularly in the Utah Chinese DLI program. The reading comprehension test consisted of ten items. See Appendix A for the external reading assessments. Although the format of the reading comprehension test was adapted to reflect the Chinese YCT test, the content was adapted to reflect text that the learners might see in the game. To reduce the priming effect on the comprehension assessments, items were randomized in both the pre- and post-

assessments. Further, although the sentence structures remained the same, the content, and thus the answers changed from pre- to post-assessments.

Teacher Informal Reading Comprehension Assessment

Two teachers who most recently taught the current class (including the current teacher) rated the students based on their reading comprehension abilities. Teachers were instructed to rate their students on a scale from 1 to 5 in terms of reading proficiency, with 5 meaning most proficient and 1 being least proficient. The teachers were told to consider student's vocabulary knowledge, oral reading fluency, and performance on in-class reading comprehension quizzes. The two scores were then correlated to check inter-rater reliability and a significant correlation was found $R = .82$ ($t = 8.52$, $df = 34$, p value < 0.001). These scores were then combined to create a composite score which was then used as a second measure for validating the scores from the stealth assessment.

Game Play Data

Finally, the source of in-game data came from the log files that capture players' in-game actions, texts read, and choices made. All players' movements and choices along with a timestamp were saved to a JSON file after each session's gameplay and then automatically uploaded to a Mongo Database each time a player saved the game. These JSON files were also stored locally after each save and then manually extracted as a backup at the end of each day. After the study was completed, all JSON files were pulled from the Mongo Database and stored into one master JSON file. This file was then wrangled into a single data frame using primarily *tidyverse* (Wickham, 2017). Appendix

D contains the *R* script with all of the libraries used to clean and analyze the data. In the following section, the specific analyses and methods that were used for each research question are addressed.

Table 4.1 provides an overview of all data collected in this study and how the data was used.

Table 4.1

Data Sources

Data source	Time	Description
Knowledge/ skills		
Vocabulary assessment	Pre- and Post-	This measures student knowledge on 45 vocabulary words that are found in the game.
Reading comprehension assessment	Pre- and Post-	There are 10 reading comprehension questions that use vocabulary and sentence structures from the game.
In-game logfiles		
Battles	In-game	This is a raw count of how many battles the student engaged in during gameplay.
Text exposure	In-game	This is a raw count of the number of texts that student was exposed to in the game.
Menu-on	In-game	This a raw account of how many times the menu was accessed during gameplay.
Time spent reading	In-game	This was a measure of time spent reading text in seconds. Time was rounded to the nearest second due to data logging complication.
Look up (vocabulary)	In-game	This counts the number of times a word was looked up in the text.
Response requested	In-game	This is a categorical variable that identifies if a text prompted a response from a participant.
Vocabulary known	In-game	This is categorical variable that identified if a vocabulary word was known, unknown, or if both in a text as determined from the pre-vocabulary test.
Duplicate text	In-game	This is a categorical variable that determined if the text was being read for the first time, one of the first three times, or more than three times.
Reading comprehension	In-game	This was a categorical variable that assigned a correct or incorrect valuable to texts that prompted learners with a question and had a definitive correct answer.

Data Analysis

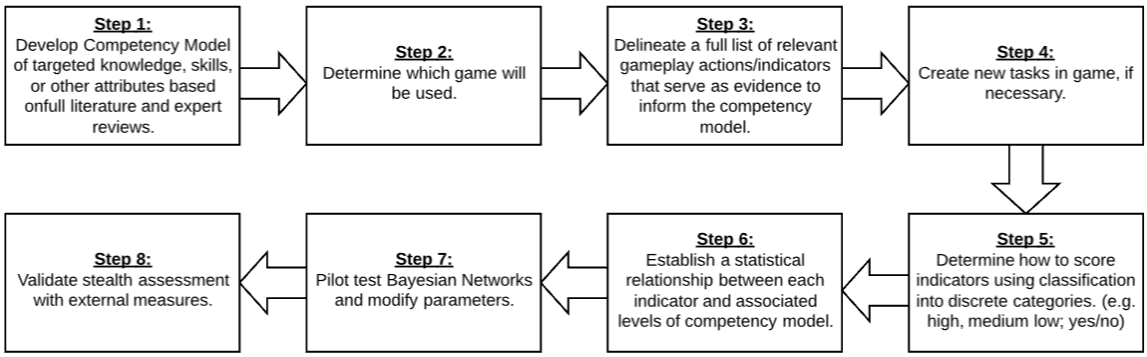
I use Bayesian Belief Networks (BBN) to analyze data from stealth assessments embedded in game play. First, I will detail the process that was used to develop the stealth assessment, and then next I will describe how the BBN was built and calibrated.

Designing Stealth Assessments

The stealth assessments in this study were designed using the eight steps outlined in Figure 4.6, which was modified from Shute et al. (2017).

Figure 4.6

Modified Approach to Designing Stealth Assessments



Step 1: Identify Key Variables. Step 1 calls for the identification of a competency model based on a literature review and expert reviews. The present study has identified Chinese reading comprehension as the skill to be assessed. Past research exploring factors that affect L2 Chinese reading comprehension have identified reading fluency, vocabulary knowledge, as well as character recognition as strong determinants for reading comprehension performance (Shen & Jiang, 2013). Further, a previous study with a similar participant sample noted the value of metacognition, and specifically self-

regulation skills while reading (Poole et al., 2018). This study found that students often used a glossing tool to check their understanding of a vocabulary word, and further that prompting that students to respond to an in-game text led to students checking their understanding of the original text. Finally, character segmentation was also identified as an important skill. However, given the tasks and game used in this study, capturing this data was not an option.

Step 2 and Step 4: Building the Task Model. These steps are concerned with selecting a game and creating tasks that elicit evidence for the desired skills. The game being used is *Legend of the Dragon,* which was described in detail in Chapter 2. This game was designed to provide an open world environment through which learners could explore and interact with NPCs with autonomy. Further, this game provided multiple L2 supports via a glossing system, battle cards with images and Pinyin support, and a quest bar to remind and direct learners towards goals. The primary tasks being assessed are learner engagement with texts within the game. The assessment draws from data collected as a result of reading these texts.

Step 3: Developing Evidence Model. This step calls for the identification of the in-game actions that will serve as indicators. Table 4.2 provides a list of the identified indicators that will be used to provide evidence for reading comprehension skills. There are a number of other variables that were left out of this table that were used in other analysis such as the number of times a player accessed an item or the menu or the amount of time spent on a particular map in the game. These were not used for this analysis as they have no relationship to reading each particular text in the game. Thus, only variables

that affected how the text was read were included.

Table 4.2

In-Game Indicators

Variable	Description
Time	Time spent reading texts in the game.
Respond	Whether or not the text being read required a response.
Sent_K	The length of the sentence.
LookUp	Whether or not the glossing tool was used while reading a text.
Vocabulary	Whether or not the text contained a known vocabulary word.
Duplicates	This indicates if the text is being read for the first, for the first three times, or if it's been read for more than three times.

Step five: Transform variables. This step calls for the discretization of the indicators see Table 4.3. This is an important process that is needed to use the software Netica (*(www.norsys.com)*) to create the BBNs. Table 4.2 identifies the key indicators for reading comprehension and have illustrated how they were discretized. The indicators were identified as a combination of variables that were deemed important from the literature, as well as those variables that were seen as important from the CART analysis in Chapter 3 of this dissertation.

The first variable *time* is a continuous variable that measures the number of seconds that a player used when reading a text. At first, this variable was divided into fast, medium, and slow reading times by identifying the mean, and then one standard deviation above and below the mean as slow and fast reading times. However, this did not make sense for a few reasons. First, there were some outliers in which students used 40 seconds to read a text. These situations were clearly moments when students clicked

Table 4.3

Discrete Categories

Variable	Description	Collected format	Discrete categories
Time	Time spent reading text in the quests.	Continuous	Fast: Less than 1 second Medium: Between 2 and 6 seconds Slow: More than 6 seconds
Respond	Whether or not the text being read required a response	Categorical	Respond No Respond
Sent_K	Length of the sentence.	Continuous	Short: Less than 6 characters Medium: Between 7 and 30 characters Long: More than 30 characters
LookUp	Whether or not a word was looked up.	Continuous	No look up Look up
Vocabulary	Whether or not a text contains a known word.	Categorical	Known Unknown Both
Duplicates	How many times the same text has been read.	Continuous	Origin Less than three times More than three times
ReadComp	Outcome variable indicating if a text was read/understood.	Categorical	Read/understood Not read/understood

on a text and then got distracted; however, these times were pulling the mean up. Due to the mean being pulled up, some texts were being identified as *fast*, but in reality, they may have been average reading times. To fix this issue, first outliers were temporarily removed and then a new mean and standard deviation was calculated. Next, any reading time that was one second or less was identified as fast, or as skipping the text. Then any reading time that was more than 2 standard deviations above the mean (> 6 seconds) was marked as *slow*. These were typically moments when a learner was seeking external help

or had gotten distracted. Anything in between, so 2 to 5 seconds was seen as optimal reading time given the average length of the texts.

The second variable *Respond* was already a categorical variable. It refers to whether or not the text that was being read prompted students to respond. This was seen as an important indicator for two reasons. First, past research has shown that when students are asked to respond to a text, they are more likely to both (a) pay attention to the text, and (b) confirm their understanding of the text (Poole et al., 2018). Given that students may re-read or focus more attention on a text, it makes sense that their reading times will also be affected, and thus this variable will likely have an effect on reading times. The two categories are respond and no respond.

The next variable *Sent_K* refers to the sentence length. This variable was originally a continuous variable that counted the total number of characters in a text. K-means clustering was used to divide this variable into three categories: short, medium, and long texts. Originally, I used a similar discretization process to the one used with *time* to categorize sentence length. However, this resulted in substantially more short texts as there were several texts with only one word. These texts occurred when players were picking up items. The second approach, K-means clustering, produced more of a balanced distribution of texts between short, medium, and long lengths. Short texts were identified as having between 0 and 6 characters, medium texts were identified as having between 7 and 30 characters, and finally, long texts were identified as having more than 30 characters.

The next variable *LookUp* is already a categorical variable that indicates whether

or not the glossing system for a word was used. *Vocabulary* is also a categorical variable that indicates whether or not a text contains a word that the student knows. There are three categories *know* refers to a text containing a word the student knows, *unknown* refers to a text that contains a word that the student does not know, and *both* means there are both known and unknown words in the text. These words were identified based on answers provided on the vocabulary pre-test and thus they are not complete.

The next variable *ReadComp* is used to track reading comprehension. Some of the texts in the game have clear right and wrong answers, for these texts the variable *ReadComp* is used to track right and wrong answers. However, in the BBN, other texts without right and wrong answers are predicted using the expert driven parameters within the BBN.

Finally, the *duplicates* variable tracks how many times a student has read a text. This variable was originally continuous and was converted to categorical by first identifying the first time a text was read (*origin)*, then identifying if the text was read less than three times, and finally, if a text was read more than three times. Table 4.4 shows the frequency for each of these categories.

Step 6, 7, 8: Establish a Statistical Relationship. The last three steps are concerned with building, piloting, and validating a statistical connection between the evidence elicited and the competency model. Shute et al. (2017) recommend using Bayesian Networks for the statistical model in stealth assessments. Thus, in the following section, a brief introduction to this analysis will be provided, followed by the procedures used to build, calibrate, and validate the analysis.

Table 4.4

Frequency of Discretized Variables

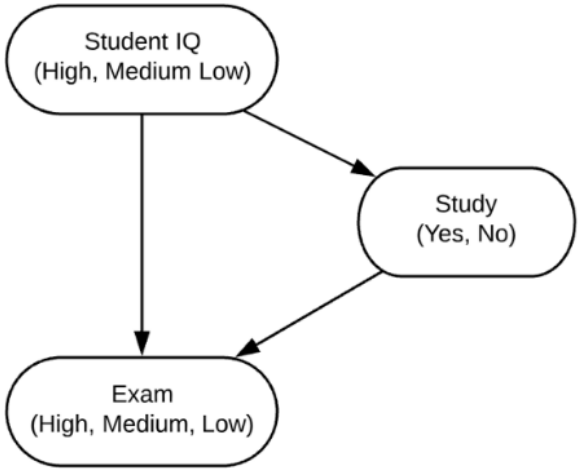
Variable	Description	Discrete categories	Frequency
Time	Time spent reading text in the quests.	Fast	12,225
		Medium	6,739
		Slow	804
Respond	Whether or not the text being read required a response	Respond	6,182
		NoRespond	13,599
Sent_K	Length of the sentence.	Short	6,513
		Medium	12,875
		Long	393
LookUp	Whether or not a word was looked up.	No Look Up	18,439
		Look Up	1,342
Vocabulary	Whether or not a text contains a known word.	Known	1,908
		Unknown	5,904
		Both	1,382
Duplicates	How many times the same text has been read.	Origin	8,153
		Less than three times	4,762
		More than three times	6,866
ReadComp	Outcome variable indicating if a text was read/understood.	Read/Understood	413
		Not Read/Understood	193

Bayesian Belief Networks

Bayesian belief networks (BBN) are statistical models that reflect how a set of variables are related by probabilities. Figure 4.7 shows a simple example of a BBN. In this example, there is a student with either a high, medium, or low IQ, a variable that determines whether the learner studies or not, and an outcome variable, exam, that is either high, medium, or low. When data is entered into the model, Bayesian networks determine the likelihood that a student with a high IQ scores low on an exam, or it can

Figure 4.7

Simple Bayesian Belief Networks Example



provide probabilities of a student with a low IQ who choose to study getting a medium score on the exam. Bayesian networks are based on Bayes Rule which states that the probability of A given B evidence is equal to the probability of B given A multiplied by the probability of B divided by the probability of A (see Figure 4.8, Note: P(A|B) = the probability of A given B occurred).

Figure 4.8

Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayesian networks are comprised of a network of conditional probability tables (CPTs) that represent nodes within a simple causal graphical structure. The nodes in the structure, also referred to as an influence diagram, illustrate the direction of influence

each node has on other nodes. In the example above, the exam score is influenced by both studying and student IQ and whether or not a student studies is further influenced by their IQ. The CPTs provide probabilities in regards to the belief that a particular state will be present given the evidence (Chen & Pollino, 2012). For instance, given the evidence that a student studies, and they have a high IQ, the CPT for Exam score might state that there is an 85% chance for a high score, 10% chance for a medium score, and 5% chance for a low score. Evidence is then entered into the model and probabilities are propagated throughout each node. For example, if a student with a high IQ who studied received a medium score, the final CPT for Exam score will update to reflect this new information.

The graphical structure of Bayesian belief networks and recent BBN software reduce the technical knowledge previously needed to build networks and make interpreting BBNs easier. In this study, the software *Netica (www.norsys.com)* will be used to construct the BBNs. Although *Netica* makes the technical side of creating BBNs easier, the conceptual development, calibration, and validation of BBNs still requires rigorous detail and documentation. Marcot et al. (2006) provide a detailed set of guidelines for building and validating BBNs which is used in this study. Speaking broadly, there are three steps involved: establishing the structure of the BBNs, defining the parameters, and then testing/validating the model.

Marcot et al. (2006) breaks the process down into three stages (alpha, beta, and gamma level), with the beta level serving as a point of reflection and opportunity for iteration on the alpha and gamma levels as the point of testing and validation of the BBN. The first step in the alpha level is to identify a structure. The BBN structure can either be

defined by an expert via literature review or can be determined through data mining approaches via the data. In most cases the expert-defined model is preferred as the data-induced model can lead to overfitting with results that make little sense (Clark, 2003; Marcot et al., 2006).

The first step is to create an influence diagram that illustrates the effect that each node has on other nodes and clearly state the rationale for each node. Figure 4.9 shows the expert-derived model, while Figure 4.10 shows the data-driven model. The data-driven model was created to illustrate a difference between the two approaches. In the data-driven model, reading comprehension is said to influence all variables, further

Figure 4.9

Expert Model

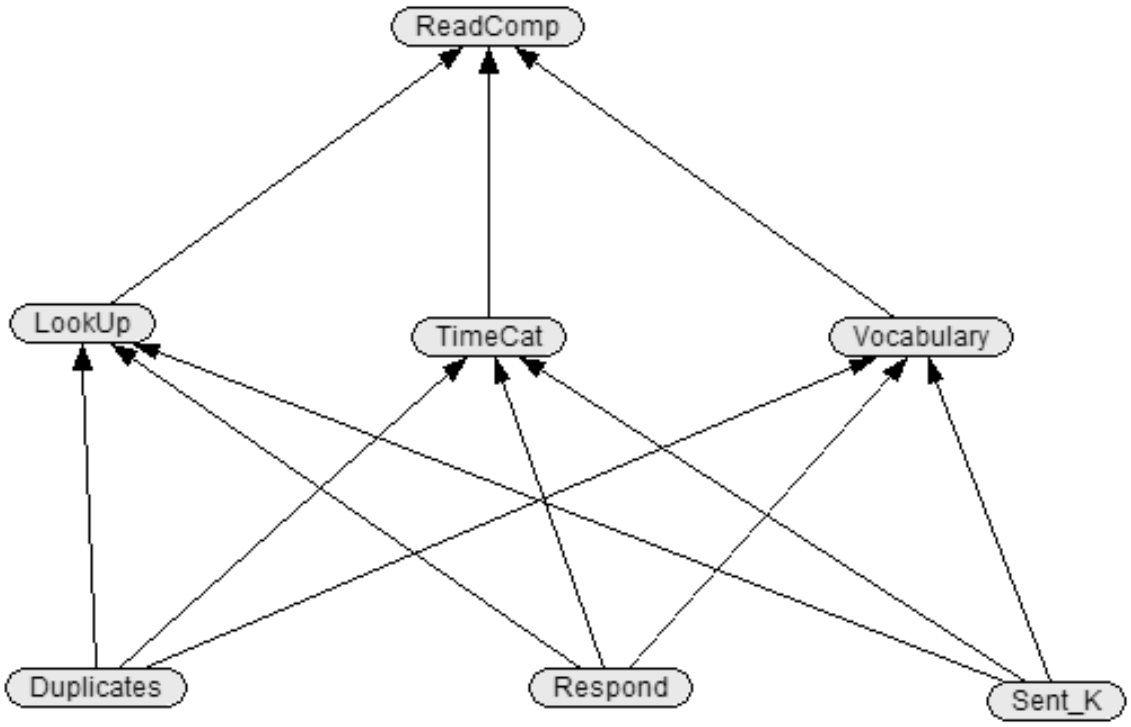
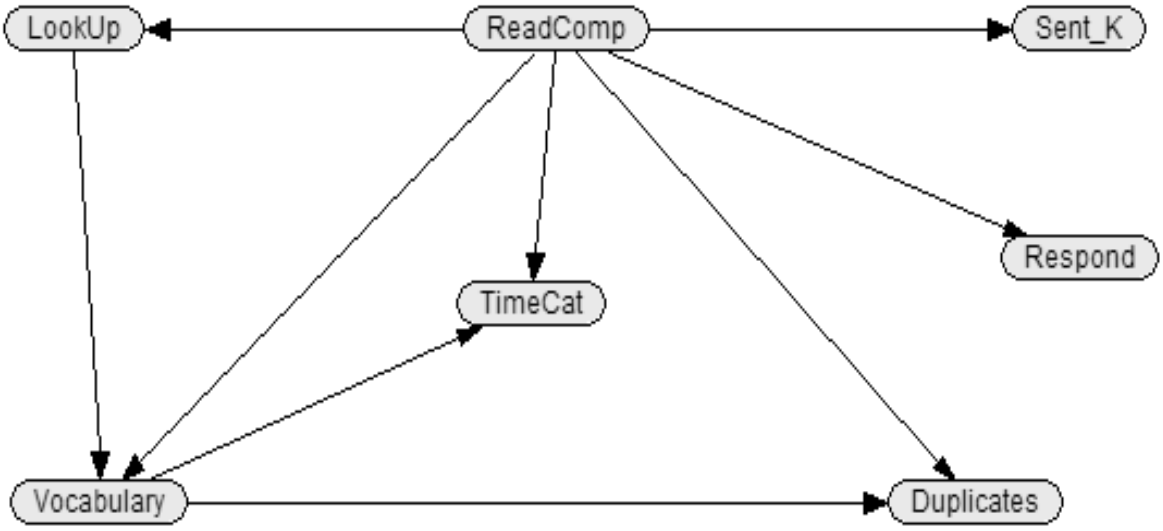


Figure 4.10

Data-Driven Model



vocabulary is shown to influence *time* and *duplicates*, while *LookUp* influences vocabulary. In the expert-driven model, only *LookUp*, *Time*, and *Vocabulary* are seen to have a direct effect on reading comprehension. *Duplicates* is seen to affect *LookUp* as the more times a learner sees a similar text, the less likely they are to look up a word. Similarly, *Duplicates* influences *Time* as repeat texts occur the amount of time spent on a text should decrease. Last, the more times a learner is exposed to a text, the more *Vocabulary* they will be exposed to. *Respond* refers to a text that prompts a learner to respond to a question. Thus, *LookUp* should be more prominent, *Time* should be longer, and *Vocabulary* should be increased given the increased amount of text with extra prompts. Similarly, *Sent_k* (the length of the text) should affect reading *Time* and *Vocabulary* given that longer texts will take longer to read and have more vocabulary. It was also assumed that longer texts may impact a learner's wish for support via the

glossing feature, *LookUp*.

The next step involves adding CPTs to each of the nodes and discretizing the variables. Marcot et al. (2006) provide a list of suggestions for this step (see Table 4.5). Many of the suggestions provided are done so to keep the parameterization process manageable. For example, the first guidelines suggest keeping parent nodes, which are nodes that influence a particular node to three or fewer and the fifth guideline suggests to keep variable categories as small as possible.

Table 4.5

Guideline for Model Construction

Guidelines for model construction
1. Number of parent nodes should be kept to three or fewer.
2. Parentless nodes should be those that have observable cases.
3. Intermediate nodes summarize major themes.
4. To the extent possible all nodes should be observable and quantifiable, in cases where they cannot the nodes should be carefully documented and explained.
5. Fewest discrete states possible, but enough to get precision
6. Number of layers should be kept to four or fewer.
7. Model should be full documented with the rationale for each node.

Note. Adapted from Marco et al. (2006).

Recall the earlier example that predicts exam score from student IQ and whether or not they studied. In this example, Exam score only has two parent nodes (study and IQ), and is categorized into three states (High, Medium, Low). Table 4.6 shows what a CPT might look like for Exam score with only two parents. It's important to note here that given that student IQ and Study are parentless nodes, they are populated with cases, or observed

data. However, the percentage, or the probability that a student with a high IQ and that studies gets a high exam score (0.85) is determined by the expert. Thus, if Exam Score were to have say four or five parents, it is easy to see how it would be difficult to track and define probabilities that make sense across all possible scenarios. In the present study, all variables were discretized according to the methods outlined above and nodes either contain two or three states (see Table 4.6). Further, all nodes have three parents, other than the three parentless nodes.

Table 4.6

Simple Conditional Probability Table

Exam score				
High	Medium	Low	Student IQ	Study
0.85	0.10	0.05	High	Yes
0.80	0.15	0.05	High	No
0.45	0.35	0.20	Medium	Yes
0.20	0.35	0.45	Medium	No
0.05	0.15	0.80	Low	Yes
0.05	0.10	0.85	Low	No

The next part involves creating *a priori* CPTs for each of the nodes. This is an initial set of probabilities that is believed to represent the relationship between each of the nodes. It is important to note here that once case data is added, the CPTs will adjust according to the data entered. Marcot et al. (2006) provide a few strategies for creating expert-defined CPTs when past data is not available as is the case in this study. One can set the CPTs to uniform value so the probabilities are set so that each state has an equal

chance of occurring. “Peg the corners” is another strategy in which the extreme cases are first defined (usually as 0% or 100%) and then the intermediate cases are set by adjusting all of the cases in between accordingly (Marcot et al., 2006). The *Peg the Corners strategy* was used to define the Reading comprehension node which has *LookUp, TimeCat,* and *Vocabulary* as parent nodes. This was seen as important to establish because the Reading comprehension knowledge has limited cases to validate and thus the outcomes will rely largely on the expert-defined CPTs. The other six nodes were set to uniform values. Although adding expert input to set CPTs is typically recommended, this was not seen as a major issue given the large number of cases available for each of these nodes. Thus, once the cases are added, the CPTs will calibrate and reflect more closely the reality of the class and student being modeled.

Table 4.7 shows the CPT that was created for the Reading Comprehension node. To define this table, I first started with all of the *LookUp* equals *No* states. I then compared vocabulary knowledge within each *Time* state. Thus, not knowing a vocabulary word and not using the glossing system to look up the unknown word was set as the lowest probability at 5% chance of reading comprehension. I did not make this impossible as there is a small chance that a word was learned via images and had been seen multiple times and thus there is still a chance that it was read. Then, for this subgroup (*NoLookUp* and *Fast* Time), I made knowing a vocabulary word the highest probability for reading comprehension at 20% and both knowing and not knowing a word as an intermediate value 10%. Next, I adjusted all of these scores for the *NoLookUp* with *medium* and *slow* reading times making the medium times the most likely to be examples

of reading and the slow reading times to be the intermediate (between fast and medium) in terms of reading comprehension probability. Finally, I copied the probabilities and pasted them into the *LookUp* equals *Yes* states and uniformly increased the probabilities of reading comprehension to reflect that using the glossing system is a strong indicator that the student is attempting to read the text. After adding all of the CPTs and then compiling the BBN, the following network was produced (see Figure 4.11). Note that all of the parents of Reading Comprehension and

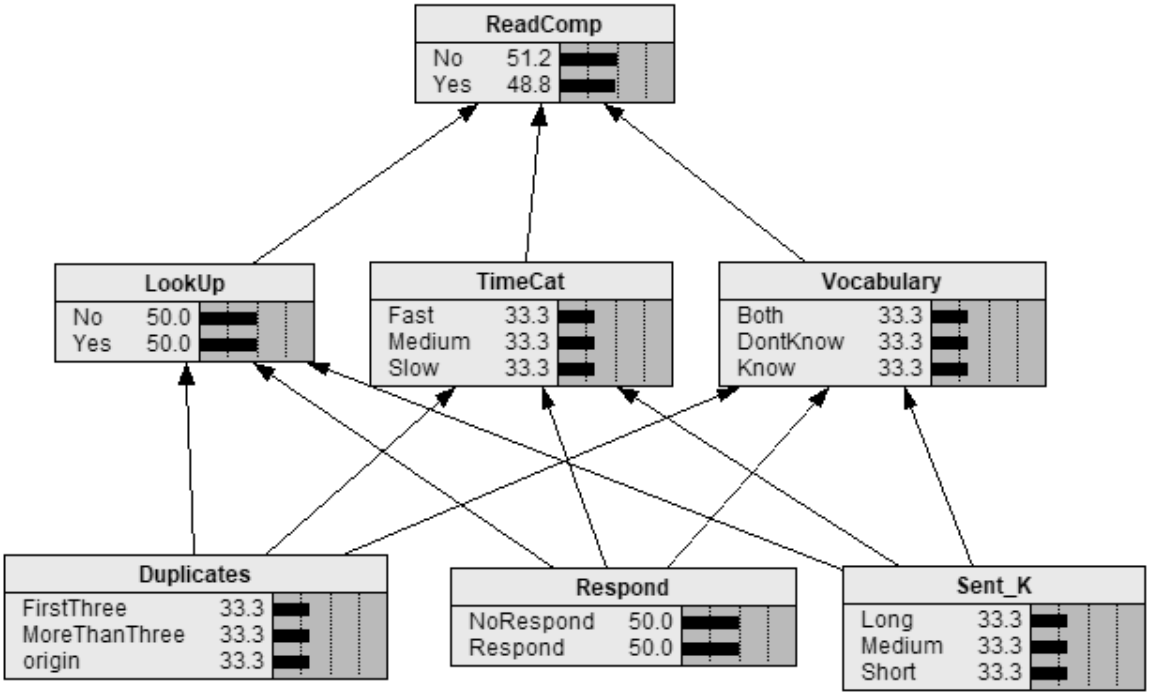
Table 4.7

Reading Comprehension Conditional Probability Table

Reading comprehension				
No	Yes	LookUp	Time	Vocabulary
90	10	No	Fast	Both
95	5	No	Fast	Don't know
80	20	No	Fast	Know
45	55	No	Medium	Both
55	45	No	Medium	Don't know
40	60	No	Medium	Know
70	30	No	Slow	Both
75	25	No	Slow	Don't know
60	40	No	Slow	Know
65	35	Yes	Fast	Both
75	25	Yes	Fast	Don't know
60	40	Yes	Fast	Know
25	75	Yes	Medium	Both
30	70	Yes	Medium	Don't know
20	80	Yes	Medium	Know
40	60	Yes	Slow	Both
45	55	Yes	Slow	Don't know
35	65	Yes	Slow	Know

Figure 4.11

Compiled Bayesian Belief Networks



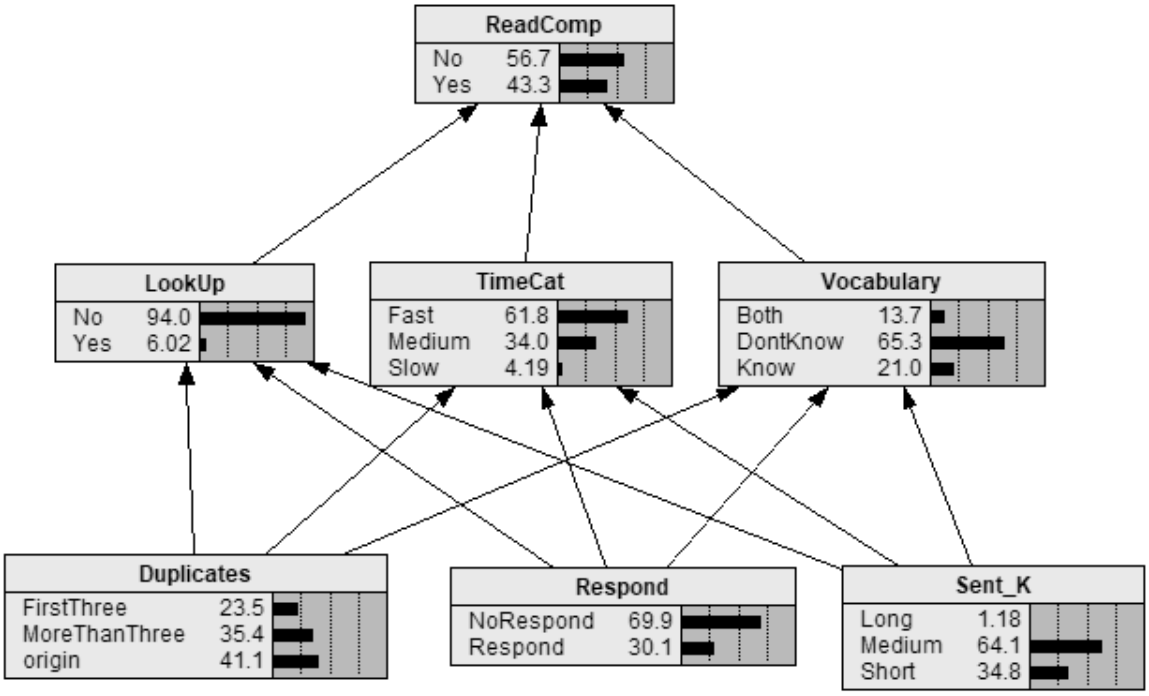
the parentless nodes are set to the probabilities that give each state an equal probability of occurring. Reading Comprehension is set so that there is a 48.8% chance that a text is understood. However, this is simply the current prediction without adding any cases to update the CPTs.

The next step involves adding cases to automatically update the CPTs. Eighty percent of gameplay data from the entire class was added to the model and the updated CPTs are reflected in Figure 4.12. Next, it is important to test and validate the network. One way to validate the network is to add evidence to a model. This involves assuming that one of the states in a node has a 100% chance of occurring and then exploring how it changes the other CPTs. In *Netica,* one can simply click on the node and state that is

desired to be 100%. For example, if *Medium* time is clicked on, it is assumed that Reading comprehension will increase, similarly with using the glossing tool, or looking up a word. To confirm that this network was working properly, I added evidence to each of the parentless nodes. Specifically, I checked *Origin, Respond,* and *Medium* all of which I assumed would lead to higher probabilities in reading comprehension for previously stated reasons. Marcot et al. (2006) state that a primary "goal is to get the model to tell you what you think it should tell you" (p. 3067). This assumption was confirmed, which led me to the next testing method.

Figure 4.12

Cases Added



Another testing method is the use of a confusion matrix to see how well the model is performing. Twenty percent of the cases from gameplay data were withheld when the

BBN was updated with cases. These 20% were then used as test data to determine the accuracy of the model to predict reading comprehension. It is important to note that overall, only 3% of the text contained prompts that had a definitive right or wrong answer. It was these texts that were used to test the accuracy of the model. Table 4.8 shows a confusion matrix of the prediction accuracy.

Table 4.8

Confusion Matrix

Predicted reading comprehension		
No	Yes	Actual reading comprehension
9	21	No
21	97	Yes

This model had an error rate of 28.3%. The model shows that it was more difficult to predict when reading comprehension was not occurring ("No") than when it was ("Yes"). Further it should be noted that there were much fewer occurrences of "Yes" than "No." Though it should be noted that behavior on the texts that have definitive answers were likely different than that of other texts. In other words, texts with definitive right and wrong answers prompted students to respond to questions and thus students likely spent the optimal amount of time on these texts and used the glossing tool. However, they still may have gotten the answer wrong. Thus, this model is looking for behavior that mimics reading comprehension rather whether or not they answered the correct answer. To further validate the model, the BBN will be applied to each students' gameplay data individually. The CPTs from the whole class BBN will be applied as the *a priori* CPTs

for each student's BBN, and then student case data will be added to the model to automatically update the competency model CPTs. Finally, the predicted probability for comprehending texts read for each student will then be correlated with student scores on the external reading assessment and informative teacher assessments of reading scores to validate the model's effectiveness to predict reading comprehension. The overall class model, and the results from each of the competency models will be discussed in the following results section.

Results

The present study was primarily focused on evaluating stealth assessments' ability to assess young Chinese dual language immersion students' reading comprehension while playing a game. To answer this question, a Bayesian Belief network was constructed using the following variables: Reading comprehension (*ReadComp*), Vocabulary look up (*LookUp*), Time (*TimeCat*), student vocabulary knowledge (*Vocabulary*), length of the text (*Sent_K*), a binary variable indicating if the text prompted learners with a response (*Respond*), and a variable that indicated how many times the text has been read (*Duplicates*). Table 4.9 describes these variables. The BBN (see Figure 4.13) illustrates the overall BBN for all students. Suggesting that 43.3% of the texts were read while 56.7% were not. A majority of those that were not read were likely skipped because they were duplicates or they were skipped because the student wanted to engage in battle. In the second row, the *LookUp* variable shows that the *LookUp* function was only used on 6% of the texts. Approximately 34% of the texts were read with the optimal amount of

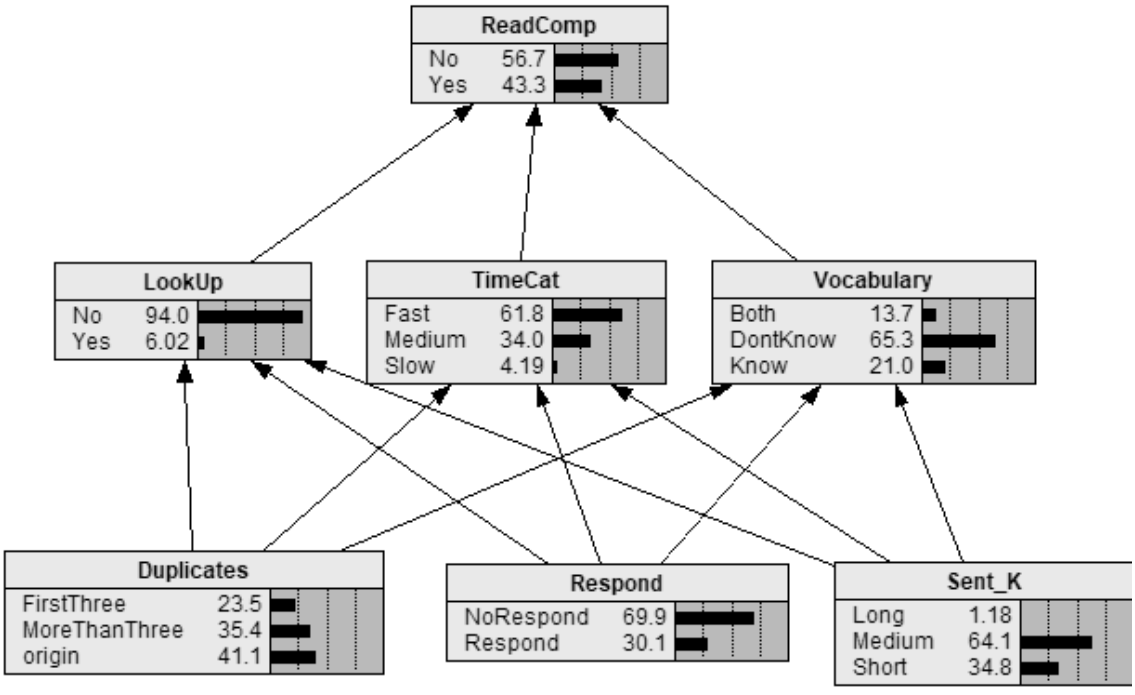
Table 4.9

Bayesian Net Variables

In-game indicators	Description
Time	Time spent reading text in the quests
Respond	Whether or not the text being read required a response
Sent_K	Length of the sentence divided into short, medium and long. K-clusters were used to divided the sentence lengths into three categories.
LookUp	This is a binary variable that indicates if a word was looked up or not in a text
Vocabulary	This is a categorical variable that indicates if a known or unknown vocabulary word or both is present in the text being read.
ReadComp	This is a binary variable that indicates if the passage was read.
Duplicates	This is categorical variable that indicates if the current text is the first time the player has seen the text, if it is one of the first three times, or if the player has seen it more than three times.

Figure 4.13

Overall Bayesian Belief Networks

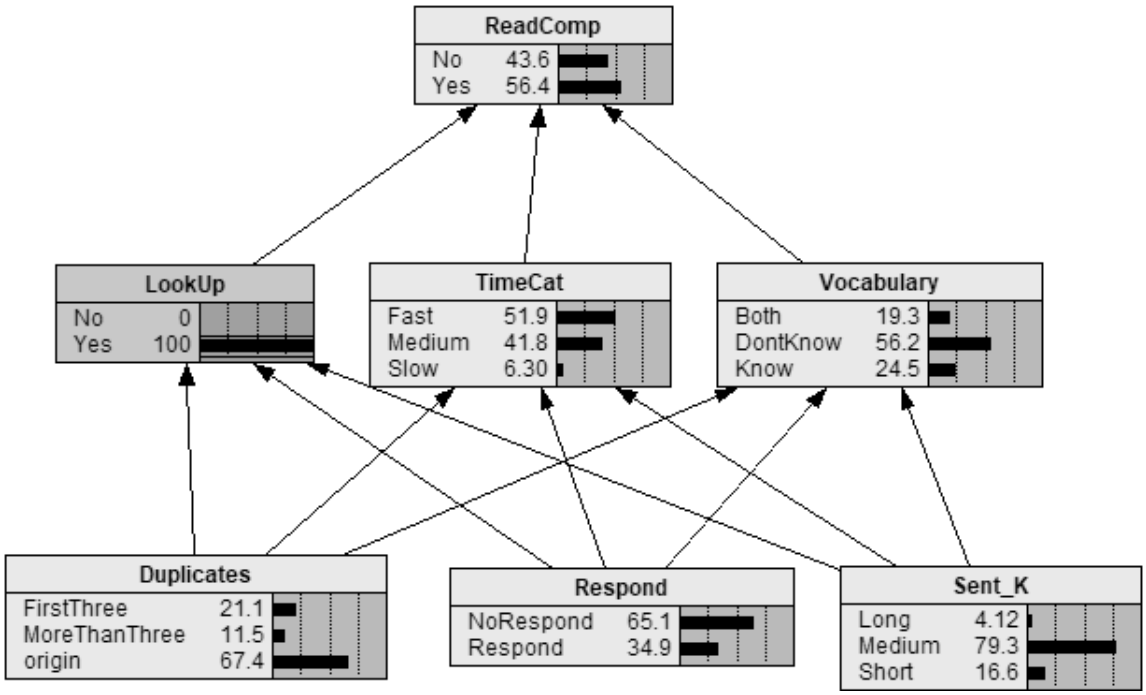


time, and 65% of the text exposed to students contained a word that they did not know. This information, in combination with the small usage rate of the look up tool, may provide rationale for the low reading rate. In the final row, 41% of the texts were original, meaning nearly 60% of the texts read were repeats. Only 30% of the texts asked students to respond and most of the texts (64.1%) were of medium length.

Figure 4.14 shows that when students did look up a vocabulary word, their chances of reading a text increase to 56.4%. Given that this is an expert model and as such, reading comprehension is defined by the expert, this finding makes sense. Note, the origin text increases to 67.4% suggesting that students are more likely to use the look up tool when they are reading a text for the first time. In addition, fast texts, or texts that are

Figure 4.14

Adding LookUp as Evidence



skipped are also reduced suggesting that when students look up a word they are not skipping it. This is an important finding for the validity of these nets. One means of validating a BBN is ensuring that it behaves as is intended (Marcot et al., 2006).

In the next figure (see Figure 4.15), the texts that result in prompting the learner with a response increases the reading comprehension from 43.3% to 49.3%. This is another example of the BBN providing evidence that it is functioning as expected.

Figure 4.15

Adding Respond as Evidence

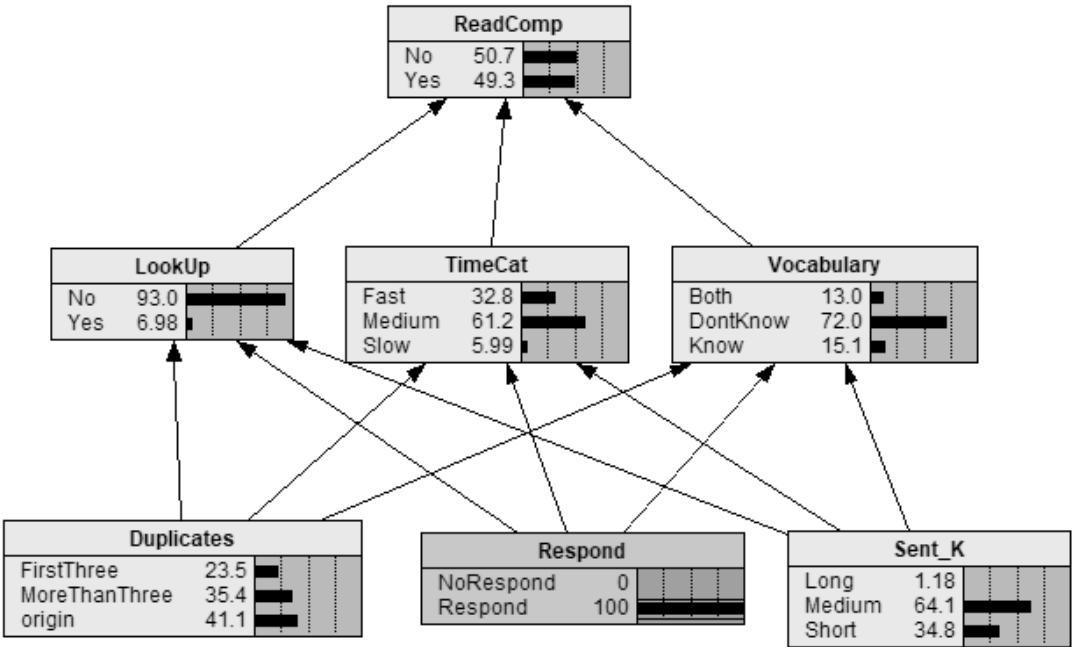
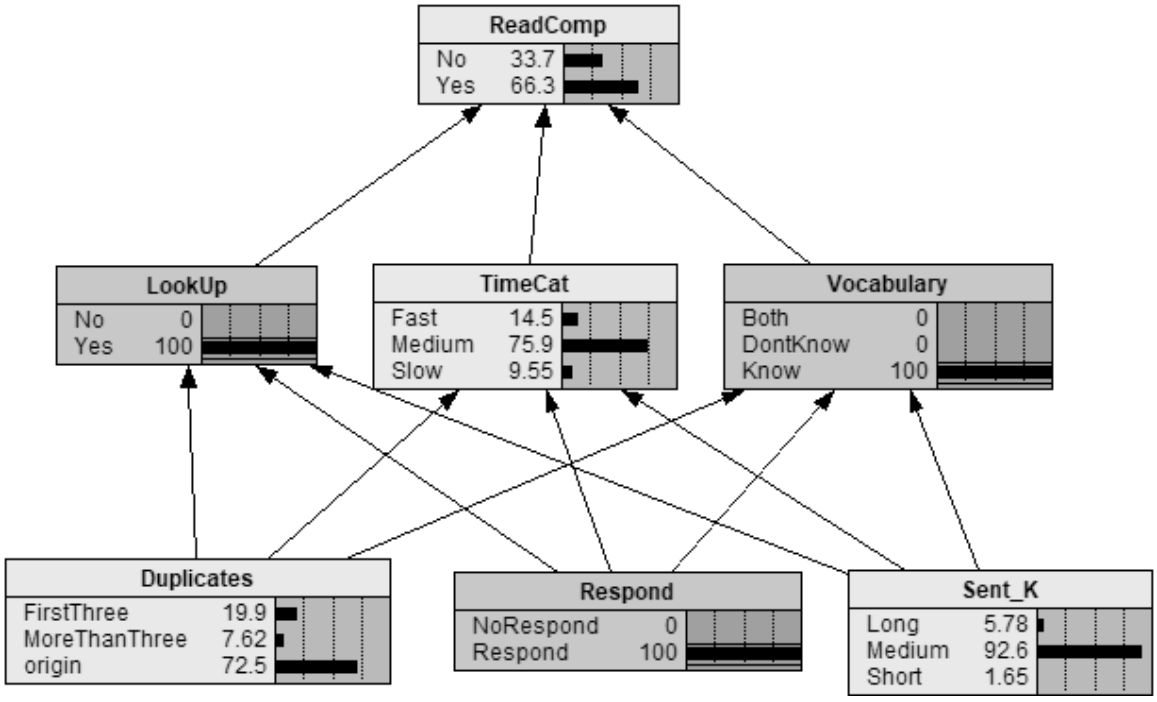


Figure 4.16 shows that when a student looks up a word, knows a vocabulary word in the text, and is responding to a prompt, there is a 66.3% chance that the student reads the text. When this happened, there was a 75.9% chance that a student used the optimal reading time (Medium) compared to only a 34% chance when reading any text. Finally, this had a 72.5% probability of occurring the first time that a student saw a text.

Figure 4.16

Adding Multiple Evidence



After exploring these models at the class level, a BBN was modeled for each student using the parameters from the overall BBN as the a priori parameters for each student. Then, student cases from the game were added to update the CPTs and the reading comprehension scores were calculated for each student based on their in-game behaviors and actions. These scores were then correlated with both their post-scores on the external reading comprehension assessment and with the informal rubric scored and averaged by two of the students' Chinese teachers. The two informal scores made by the Chinese teachers were significantly correlated $R = .82$ ($t = 8.52$, $df = 34$, p value < 0.001), suggesting that the teacher score has high reliability. The BBN reading comprehension scores for the students was significantly correlated with the external reading

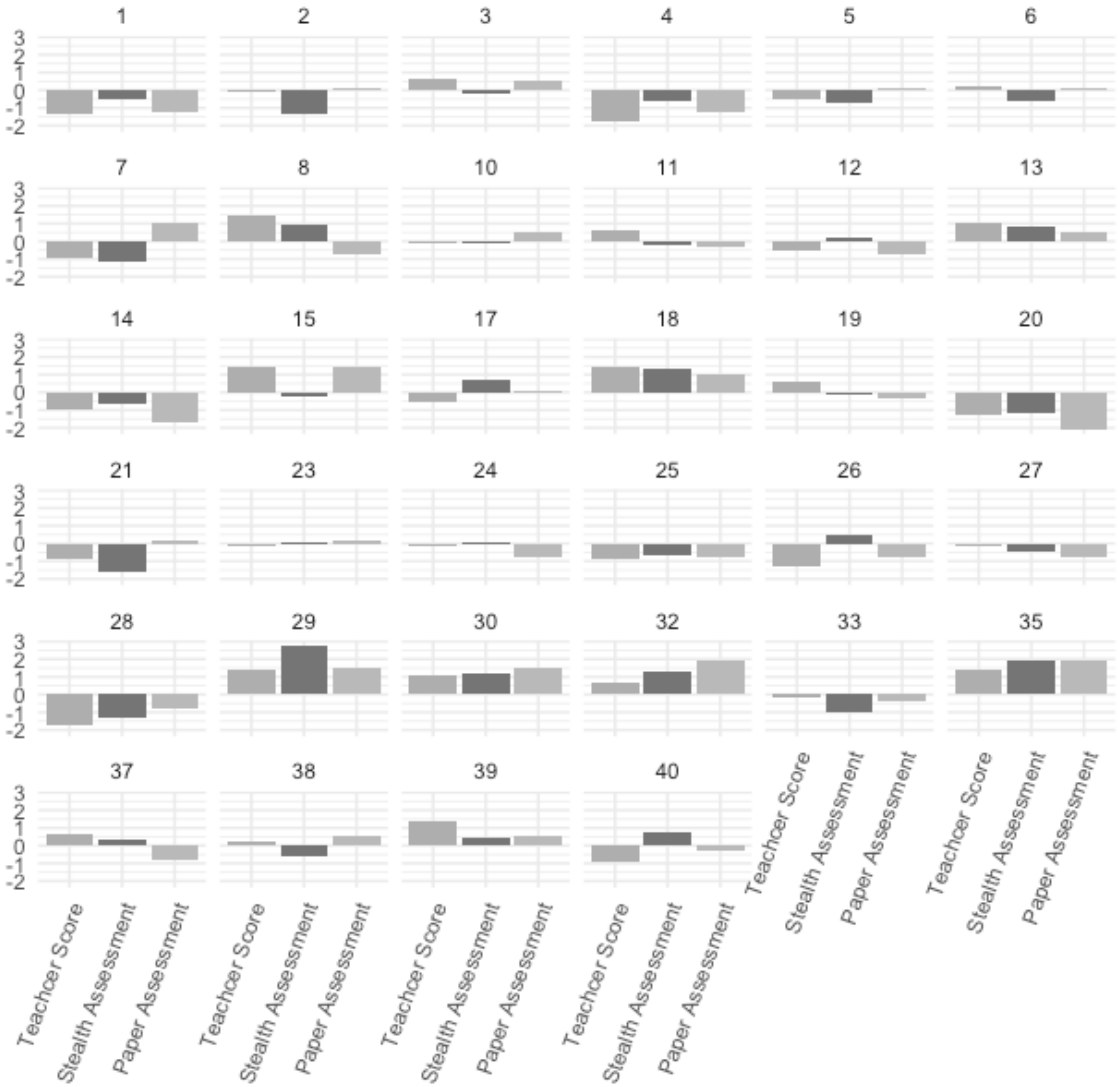
comprehension test $R = .52$ ($t = 3.45$ df = 32, p value = 0.001). Further, there was a stronger correlation with the teacher scores $R = .65$ ($t = 4.93$, $df = 34$, p value < 0.001).

To visualize these correlations, a facet-grid plot was produced (see Figure 4.17) in *R* to illustrate the BBN score, teacher score, and external post-assessment score for each student. Scores were first standardized so that 0 is equal to the average score for that class. Bars above 0 represent a score that was above the average for the given assessment and likewise scores below 0 represent scores that were below the average for that particular assessment. Values on the Y axis represent standard deviations from the mean. Ideally, scores will be similar across all three assessments to lend validity to the BBN as an assessment. For example, student 35 is approximately 2 standard deviations higher than the mean on all three assessments, student 23 is at the mean on all three assessments, and student 20 is 1 to 2 standard deviations below the mean all three measures. This visualization also helps illustrate areas for concern, for example Student 33 and 6 are about 1 standard deviation below the mean on the BBN assessment, but average or higher on the other two assessments.

Finally, to further explore how the stealth assessments compare to the two external measures, the final reading competency scores produced by the Bayesian belief networks were discretized into three categories: low, medium, and high. Low scores (n = 8) were those that were 1 standard deviation below the mean. High scores (n = 9) were scores that were 1 standard deviation above the mean. Medium scores (n = 19) were those within 1 standard deviation of the mean. Then, the average paper-based and teacher score for students in each of the stealth assessment categories (e.g., low, medium, high) were

Figure 4.17

Student Scores on All Three Assessments



calculated. Table 4.10 shows the results. An ANOVA confirms that students in the stealth assessment categories have significantly different scores on the paper-based assessments [F(2,31) = 5.08, p = 0.01] with a large effect size (Cohen's f = .57) and on the teacher scores [F(2,33) = 7.35, p = .002] also with a large effect size (Cohen's f = .67). A Tukey

HSD post-hoc test indicated that students in the High category of stealth assessments have significantly higher means on teacher scores than students in the medium (p = .03) and low (p = .002) categories. Students in the High category of stealth assessments have significantly higher means on the paper-based scores than students in the medium (p = .01) category, but not in the low (p = .05) category. These implications for these findings will be further explored in the discussion section.

Table 4.10

Comparing Stealth Assessments to External Measures

	Stealth assessments						
	Low		Medium		High		
Assessments	*M*	*SD*	*M*	*SD*	*M*	*SD*	ANOVA *F* value
Paper-based assessment (Range = 1-10)	4.00	2.36	4.11	1.79	6.55	2.18	5.08*
Teacher score (Range = 1-5)	2.19	0.75	3.05	1.14	4.17	1.15	7.35**

* $p < .01$.

** $p < .001$.

Discussion

To date, no studies have explored the development and implementation of game-based assessments in an L2 context. Assessing learners while they play a digital game that promotes learning and meaningful interaction could be especially meaningful for dual language immersion (DLI) programs that must teach both language skills and content knowledge in a limited time. The present study explored if stealth assessments in

a digital game could effectively assess L2 Chinese students' reading comprehension skills. To do this, Bayesian belief networks were built to model reading that occurred by each student while playing the game. The parameters were first established from conditional probability tables that were created by using class-level data as a priori parameters. Student data was then entered into the model to update beliefs about reading comprehension for each student. These scores were then extracted and correlated with both the external post-reading assessment and informal assessments of each student by their two most recent Chinese teachers. The results showed that the scores produced by the BBN were significantly correlated (R = .52) with the paper-based post-external assessment. This suggests that by leveraging student interaction with the glossing tool, average time spent on task, and vocabulary knowledge, the stealth assessments were predictors of traditional reading assessment measures. This is similar to that of Shute et al. (2016) who found a significant correlation of R = .43 while modeling problem solving that occurred within the commercial game *Plants vs Zombies*. Shute et al. compared their stealth assessment scores to two external measures and concluded that the MicroDyn problem solving assessment (Wüstenberg et al., 2012) was a better predictor of the problem-solving scores from the stealth assessments in *Plants vs Zombies* because the tasks were more similar. Similarly, I also correlated reading comprehension scores found in the stealth assessments with a second external measure. Teachers informal scores were found to have a stronger significant correlation (R = .66). While the teacher scores are admittedly based on intuitive assumptions made by the educators, it is important to note that these two teachers' scores were highly significantly correlated (R = .82) and were

likely based on several experiences (not a one-off snapshot) with the students attempting to read in the classroom. Both of these teachers reported delivering a weekly, formative reading assessment with each of their students. Thus, this may be evidence that the BBN is acting closer to an educator's informal assessment in that it takes into account more data when predicting the probability that a student comprehends a text.

In a follow-up analysis, I created three categories of students (low, medium, and high) based on the results of the stealth assessments and then compared how these three categories scored on the paper-based assessments and the teacher scores. Given that the students who scored in the high category in the stealth assessment had significantly higher scores than those that scored in the medium category, this would suggest that the stealth assessments are good at identifying highly proficient readers. Differences between students in the low and medium categories were not significant. This may suggest that the stealth assessment does not distinguish between lower- and medium-level readers as well as highly proficient readers. However, this could also be a result of my classification of proficiency based on the stealth assessment scores.

To further understand the implications of these findings, it is important to understand what each assessment is capturing. In other words, teacher scores are likely a result of teacher intuition based on multiple observations of students' performance on informal reading assessments in the classroom. Whereas, the paper-based assessment is a one-time snapshot of a student's ability to extract information from a text. The stealth assessment is most likely capturing a student's engagement and/or attempts to read a text. In other words, the stealth assessment is not capturing whether or not a student has

accurately extracted information, but rather if a student is actively reading a text. It is reasonable to assume that these three outcome variables (i.e., the stealth assessment, pre- and post-assessment, and teacher assessment) would be highly correlated given that to extract information, one must actively read. Thus, stealth assessments may be better at identifying more proficient readers simply because they collect more observations and given their proficiency, they are able to engage with the text more. While these findings are encouraging, more research is needed in order for stealth assessments to be used more broadly in L2 contexts.

Conclusion

The goal of the present study was to build and evaluate stealth assessments to assess L2 Chinese reading comprehension while they played a digital game in the classroom. Assessing L2 reading comprehension is difficult because the process is internal to the learner (J. Lee & VanPatten, 2003) and reading depends on many factors including vocabulary knowledge, grammar knowledge, L1 reading abilities, and among others (Jeon & Yamashita, 2014). Game-based assessments may help address some of the difficulties associated with assessing L2 reading comprehension. Through game-based assessments, researchers can not only collect large amounts of data related to the text that students read and how students interact with the texts, but they can also include data collected outside of the game to further strengthen reading models. The present study lays the ground work for building such models by illustrating how average time spent on text, use of an in-game glossing tool, and vocabulary knowledge taken from a pre-test are associated with reading proficiency measured by both a paper-based test and informal

teacher assessments.

Implications and Limitations

The results of the present study have several implications for future research. First, if games can continue to accurately predict students' reading skills, then games can be further designed to adapt to a learner's reading level as they improve. Thus, as a student plays the game, becomes exposed to more vocabulary, and develops reading skills, new levels can be opened that allow the learner to continue to be challenged and exposed to new vocabulary and text. Second, the use of BBNs in the present study not only illustrate how such networks can be built and applied to assessments, but they also provide user-friendly visuals that may be beneficial to educators. Future research should explore ways of providing this information to teachers and how teachers can use this information to inform practice. Third, the present study was limited by the amount of vocabulary knowledge that was able to be considered. In other words, I was only able to assess 45 words on the pre- and post-assessments. To build better models, I'll need more information about each of the characters that students know. Accuracy could be improved by continuously monitoring student vocabulary knowledge throughout the year, through a mobile app, and then adding that vocabulary knowledge to the BBNs. This would allow more accurate vocabulary states in which the model would know if all of the vocabulary words were known rather than if only the target vocabulary word was known. Next, the BBNs illustrated the value of understanding how duplicate texts affect reading comprehension and further how prompting learners to respond to a text promotes reading comprehension. This does not mean that every text should prompt a learner to respond as

this may disrupt the flow of gameplay, but designers should be aware that such prompts will increase learner attention. Finally, the present study only used one type of text that had a definitive response. This does not mean that more texts with definitive responses should be added as this may again lead learners to feel that they are being assessed or playing an educational game, but more unique text types should be found that can be integrated into the dialogue to capture more unique types of reading to be assessed.

The present study has a few limitations. First, the sample size is relatively small for this type of study. Ideally, the parameters that were used for the student models would have come from a larger population instead of the classes that was being observed. This is important because the model was assessing the learners based on their relative performance compared to their classmates. The students in this school may have played the game in a particular way that is different from learners of the same age group at large. Future studies will need to expand these assessments to more classrooms. Another limitation was in the lack of vocabulary knowledge and task type in the game. The lack of vocabulary knowledge makes it difficult to fully understand how much of a text that a student knows. The BBNs are good at making inferences from this lack of data, but more data would undoubtedly increase the accuracy of the models. Task type diversity is also important. Future studies will need to explore reading comprehension when students are engaging in a wider variety of tasks. Further, a wider variety of tasks will provide more information to the BBN, which can also lead to models with better accuracy.

CHAPTER 5

DESIGNING A DIGITAL GAME FOR CLASSROOM USE, DATA COLLECTION AND GAME-BASED ASSESSMENT: MULTIPLE PAPER DISSERTATION

This dissertation set out to build and evaluate a digital game designed to both promote L2 learning in a Chinese DLI classroom context and assess L2 Chinese reading comprehension via game-based assessments. Although past research has identified games as beneficial sites for L2 learning and motivation (Cornillie et al., 2012; Hayes, 2005; Prensky, 2001; Warschauer & Healey, 1998), to date, few studies have explored the integration of digital games into elementary classrooms (deHaan, 2019; Jones, 2020), the use of EDM techniques to investigate how gameplay styles relate to learning, and no study has specifically designed games with the intent to assess L2 learners. In addition, very few L2 studies conduct research on learning Chinese. Further, the present study attempted to measure L2 reading, which is a skill that is difficult to capture given that many of the processes involved are internal to the learner (J. Lee & VanPatten, 2003).

Research has identified several benefits associated with using virtual assessments, and, more specifically, game-based assessments. Such benefits include providing learners an opportunity to learn while being assessed. This is possible because unlike traditional assessments in which authentic L2 use is often divorced from the assessment items, the assessments in GBA are embedded into a context that is meaningful and familiar to learners (Clarke-Midura & Dede, 2010; Halverson & Owen, 2014). Another benefit is reduced anxiety. Research indicates test-taker anxiety has a negative effect on test

performance (Cakici, 2016; Gardner et al., 1997). Given that stealth assessments are 'stealthy' and unobtrusive, learners are often unaware that they are being assessed, thus reducing anxiety associated with tests (Mavridis & Tsiatsos, 2017). Finally, given the fun factor that is typically associated with games (Ansteeg, 2015; Becker, 2007; Gee, 2003; Prensky, 2001), and the amount of time learners may potentially spend on games, another added benefit is the amount of data that can be collected. Traditional assessments have been argued to be problematic because they take a snapshot approach to the skill being assessed (Baker et al., 2016; Shute & Wang, 2015). However, because game-based assessments can collect multiple observations of students' performance and utilize data collected at multiple time points, they may be able to provide more valid assessments of a learner's skill.

Thus, the goal of this dissertation was to build a game with embedded assessments designed to assess L2 Chinese reading comprehension. In Chapter 2, I explored the learning that occurred as a result of playing the game in the classroom and found significant gains in both L2 vocabulary and reading comprehension. Further, I identified seven types of support that were provided to the students as they played the game. This study found that such supports not only help learners become better game players and provide linguistic support, but also such support provides a linguistically rich environment for learners to further use their L2.

In Chapter 3, I illustrated how educational data mining techniques could be used with data collected from a digital game in an L2 context. This study found that time spent on reading texts was an important indicator for both vocabulary and reading

comprehension gains and further that the use of the in-game glossing tool also accounted for some of the variance in learning gains. Finally, this study identified four subgroups within the two classes based on how students played the game. One group was more focused on battling, one on reading the text, another on using the glossing tool more than others, and finally one group appeared to struggle with the game. There were no significant differences in learning gains between these two groups suggesting that the regardless of style of gameplay, learners were able to learn as a result of the intervention.

Finally, in Chapter 4, variables identified in Chapter 3 were used to build a Bayesian Belief network that assessed the likelihood that students read and understood a text. A BBN was created and a final probability of reading comprehension was calculated for each student. These scores were then correlated with an external reading comprehension test and informal assessments from two past teachers. The game-based assessments were significantly correlated with both assessments, but a stronger correlation was found between the game-based assessment and the informal teacher assessments.

The results for the present study have several implications for the digital games-based language learning field. First, this study provided a model for how digital games can be integrated into classroom settings. By using supplemental material, in-game exploits can be leveraged for L2 language practice outside of the game. Further, by providing support to learners while they play the game, instructors not only provide linguistic and game literacy support, but they can also leverage the game as a means to engage in meaningful L2 interactions. Further research should explore how educators

leverage digital games in the classroom and the type of supports that are needed for other L2 gaming contexts.

Second, the EDM approaches used in this game can be used to make inferences about how the game promoted learning. In addition, the data collected and analyzed via the game could be utilized to inform teacher practices when teaching with and around games. Future research is needed to explore the effect of teacher support given in the classroom on the identified in-game indicators. In other words, it is important to know if providing support to learners made average time spent on text an important variable or if this variable would be important even when teacher support is removed. In addition, future research should explore what data is most useful to educators who use games in the classroom. In other words, would teachers leverage data that indicates how certain students play a game?

Finally, the stealth assessments were strongly correlated with external assessments and informal teacher assessments on student reading abilities. Before these assessments are used to assess learners in lieu of traditional paper-based assessments, it will be important to increase accuracy and test their reliability across grade level, languages (e.g., Spanish, German), and schools. Further, the assessments will need to be designed to assess other L2 competencies (e.g., vocabulary, speaking). If game-based assessments can be validated on a larger scale, educators can leverage games as a means to teach and assess their learners simultaneously.

In my dissertation, I set out to build a game designed to promote learning and assess L2 Chinese reading comprehension. This study found that learners significantly

increased vocabulary knowledge and reading comprehension after playing the game over seven 20-minute sessions. Further this study identified four subgroups that had unique gaming styles and how those gaming styles may have led to different learning outcomes and gaming experiences. Finally, the stealth assessments in this game were significantly correlated to both teacher assessments and paper-based assessments suggesting that this game successfully measured students' L2 reading skills.

REFERENCES

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. Routledge/Taylor & Francis.

Alyaz, Y., & Genc, Z. S. (2016). Digital game-based language learning in foreign language teacher education. *Turkish Online Journal of Distance Education, 17*(4). 130-146.

Ang, C. S., & Zaphiris, P. (2006). Developing enjoyable second language learning software tools: A computer game paradigm. In P. Zaphiris & G. Zacharia (Eds.), *User-centered computer aided language learning* (pp. 1-22). Idea Group.

Almond, R. G., Kim, Y. J., Velasquez, G., & Shute, V. J. (2014). How task features impact evidence from assessments embedded in simulations and games. *Measurement: Interdisciplinary Research & Perspectives*, *12*(1-2), 1-33.

Ansteeg, L. W. (2015). Incidental lexicon acquisition through playful interaction. *International Journal of Emerging Technologies in Learning (iJET)*, 10(1). 4-10.

Attewell, P., Monaghan, D., & Kwong, D. (2015). *Data mining for the social sciences: An introduction*. University of California Press.

Baker, R. S., & Clarke-Midura, J. (2013, June). Predicting successful inquiry learning in a virtual performance assessment for science. In *International conference on user modeling, adaptation, and personalization* (pp. 203-214). Springer.

Baker, R. S., Clarke-Midura, J., & Ocumpaugh, J. (2016). Towards general models of effective science inquiry in virtual performance assessments. *Journal of Computer Assisted Learning*, *32*(3), 267-280.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds), *Learning analytics: from research to practice* (pp. 61-75). Springer.

Baltra, A. (1990). Language learning through computer adventure games. *Simulation & Gaming: An Interdisciplinary Journal, 21,* 445-452.

Becker, K. (2007). Digital game-based learning once removed: Teaching teachers. *British Journal of Educational Technology*, *38*(3), 478-488.

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics, 25*, 133-150. doi:10.1017/S0267190505000073

Bernhardt, E. (2011). *Understanding advanced second-language reading.* Routledge.

Block, E. (1992). See how they read: Comprehension monitoring of L1 and L2 readers. *TESOL Quarterly*, 26(2), 319-343.

Bravo-Agapito, J., Bonilla, C. F., & Seoane, I. (2020). Data mining in foreign language learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(1), 1-16.

Burkhauser, S., Steele, J., Li, J., Slater, R., Bacon, M., & Miller, T. (2016). Partner language learning trajectories in dual-language immersion: Evidence from an urban district. *Foreign Language Annals*, 49(3), 415–431.

Bytheway, J. (2014). In-game culture affects learners' use of vocabulary learning strategies in massively multiplayer online role-playing games. *International Journal of Computer-Assisted Language Learning and Teaching*, *4*(4), 1-13.

Cakici, D. (2016). The correlation among EFL learners' test anxiety, foreign language anxiety and language achievement. *English Language Teaching*, *9*(8), 190-203.

Calvo-Ferrer, J. R. (2017). Educational games as stand-alone learning tools and their motivational effect on L2 vocabulary acquisition and perceived learning gains. *British Journal of Educational Technology*, *48*(2), 264-278.

Chen, H. J. H., & Yang, T. Y. C. (2013). The impact of adventure video games on foreign language learning and the perceptions of learners. *Interactive Learning Environments*, *21*(2), 129-141.

Chen, S. H., & Pollino, C. A. (2012). Good practice in Bayesian network modelling. *Environment Modelling & Software*, *37*, 134-145.

Chin, D. B., Blair, K. P., & Schwartz, D. L. (2016). Got game? A choice-based learning assessment of data literacy and visualization skills. *Technology, Knowledge and Learning*, *21*(2), 195-210.

Christian, D. (2011). Dual language education. In Hinkel, E. (Ed.), *Handbook of research in second language teaching and learning* (Vol. 2, pp. 3-20). Routledge.

Clark, J. S. (2003). Uncertainty and variability in demography and population growth: A hierarchical approach. *Ecology*, *84*(6), 1370-1381.

Clarke, M. A. (1980). The "short-circuit" hypothesis of ESL reading—or when language competence interferes with reading performance. *Modern Language Journal, 64,* 203–209.

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education*, *42*(3), 309-328.

Cobb, T., & Horst, M. (2011). Does word coach coach words. *Calico Journal*, *28*(3), 639-661.

Coleman, D. W. (2002). On foot in SIM CITY: Using SIM COPTER as the basis for an ESL writing assignment. *Simulation & Gaming*, 33(2), 217-230.

Collentine, K. (2011). Learner autonomy in a task-based 3D world and production. *Language Learning & Technology, 15*(3), 50-67.

Cornillie, F., Clarebout, G., & Desmet, P. (2012). Between learning and playing? Exploring learners' perceptions of corrective feedback in an immersive game for English pragmatics. *ReCALL*, *24*(03), 257-278.

Dalton, G., & Devitt, A. (2016). Action research Irish in a 3d world: Engaging primary school children. *Language Learning & Technology*, *20*(1). 21-33.

De Grove, F., Van Looy, J., & Courtois, C. (2010). Towards a serious game experience model: Validation, extension and adaptation of the GEQ for use in an educational context. In L. Calvi, K. C. M. Nuijten, & H. Bouwknegt (Eds.), *Playability and player experience* (Vol. 10, pp. 47-61). Breda University of Applied Sciences.

deHaan, J. (2019). Teaching language and literacy with games: What? How? Why? *Ludic Language Pedagogy* (1), 1-57.

Dew, J. E. (1994). Back to basics: Let's not lose sight of what's really important. *Journal of the Chinese Language Teachers Association*, *29*(2), 31-46.

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Educational Technology & Society, 17*(1), 17–28.

Dourda, K., Bratitsis, T., Griva, E., & Papadopoulou, P. (2014). Content and language integrated learning through an online game in primary school: A case study. *Electronic Journal of e-Learning*, *12*(3), 243-258.

Egenfeldt-Nielsen, S. (2007). *Beyond edutainment: The educational potential of computer games.* Continuum Press.

Erhel, S., & Jamet, E. (2016). The effects of goal-oriented instructions in digital game-based learning. *Interactive Learning Environments*, *2*4(8), 1744-1757.

Everson, M. E. (1998). Word recognition among learners of Chinese as a foreign language: Investigating the relationship between naming and knowing. *The Modern Language Journal*, *82*(2), 194-204.

Fan, K. Y., Gao, J. Y., & Ao, X. P. (1984). Pronunciation principles of the Chinese character and alphabetic writing scripts. *Chinese Character Reform* 3*,* 23-37.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37-54.

Feldman, L. B., & Siok, W. W. (1999). Semantic radicals in phonetic compounds: Implications for visual character recognition in Chinese. In J. Wang, A. W. Inhoff, & H.-C. Chen (Eds.), *Reading Chinese script: A cognitive analysis* (pp. 19-35). Erlbaum.

Franciosi, S. J., Yagi, J., Tomoshige, Y., & Ye, S. (2016). The effect of a simple simulation game on long-term vocabulary retention. *CALICO Journal*, *33*(3), 355-379.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI Magazine*, *13*(3), 57-70.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*, 2-18.

Formann, A. K. (1984). *Die latent-class-analyse: Einführung in die theorie und anwendung*. Beltz.

Fortune, T. W., & Tedick, D. J. (2008). One-way, two-way and indigenous immersion: A call for cross-fertilization. In T. W. Fortune & D. J. Tedick (Eds), *Pathways to multilingualism: Evolving perspectives on immersion education* (pp. 3-21). Multilingual Matters.

Fotouhi-Ghazvini, F., Earnshaw, R., Robison, D., & Excell, P. (2009). The MOBO City: A mobile game package for Technical language learning. *International Journal of Interactive Mobile Technologies*, *3*(2), 19-24.

Foung, D. (2019). Making good suggestions in analytics-based early alert systems: Shaping minds and changing behaviours. *Journal of Applied Research in Higher Education, 12*(1), pp. 109-123. https://doi.org/10.1108/JARHE-12-2018-0264

Galaup, M., Segonds, F., Lelardeux, C., & Lagarrigue, P. (2015). Mecagenius: An innovative learning game for mechanical engineering. *International Journal of Engineering Education, 31*(3), 786-797.

Gardner, R. C., Tremblay, P. F., & Masgoret, A. M. (1997). Towards a full model of second language learning: An empirical investigation. *The Modern Language Journal, 81*(3), 344-362.

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE)*, *1*(1), 1-4.

Gibson, D., & Clarke-Midura, J. (2015). Some psychometric and design implications of game-based learning analytics. In D. Ifenthaler, J. Spector, P. Isaias, & D. Sampson (Eds.), *E-learning systems, environments and approaches: Theory and implementation* (pp. 247-261). Springer.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining, 4*(1), 111-143.

Godwin-Jones, R. (2017). Scaling up and zooming in: Big data and personalization in language learning. *Language Learning & Technology*, *21*(1), 4-15.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Ernst Klett Sprachen.

Grabe, W. (2014). Key issues in L2 reading development. In *Proceedings of the 4th CELC Symposium for English Language Teachers-Selected Papers* (pp. 8-18). National University of Singapore.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. Longman.

Guan, C. Q., Liu, Y., Chan, D. H. L., Ye, F., & Perfetti, C. A. (2011). Writing strengthens orthography and alphabetic-coding strengthens phonology in learning to read Chinese. *Journal of Educational Psychology*, 103(3), 509-522.

Hadley, A. (2001). *Teaching language in context*. Heinle & Heinle.

Halverson, R., & Owen, V. E. (2014). Game-based assessment: an integrated model for capturing evidence of learning in play. *International Journal of Learning Technology, 9*(2), 111-138.

Hayes, E. B. (1988). Encoding strategies used by native and non-native readers of Chinese Mandarin. *The Modern Language Journal*, *72*(2), 188-195.

Hayes, E. (2005). Women and video gaming: Gendered identities at play. *Games & Culture*, *2*(1), 23-48.

Hinkel, E. (Ed.). (2017). *Handbook of research in second language teaching and learning* (Vol. 3). Routledge.

Hitosugi, C. I., Schmidt, M., & Hayashi, K. (2014). Digital game-based learning (DGBL) in the L2 classroom: The impact of the UN's off-the-shelf videogame, Food Force, on learner affect and vocabulary retention. *Calico Journal*, *31*(1), 19-39.

Hong, M. (2018, April). *Exploratory data mining with classification and regression trees (CART): An introduction to CART*. American Psychological Association. https://www.apa.org/science/about/psa/2018/04/classification-regression-trees

Hong, W. (1997). Multimedia computer-assisted reading in business Chinese. *Foreign Language Annals,* 30(3), 335-344.

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *Canadian Modern Language Review, 61*(3), 355-382.

Hsiao, I. Y. T., Lan, Y.-J., & Kao C.-L., & Li, P. (2017). Visualization analytics for second language vocabulary learning in virtual worlds. *Educational Technology & Society*, 20 (2), 161–175.

Hsu, H. Y., & Wang, S. K. (2010). Using gaming literacies to cultivate new literacies. *Simulation & Gaming*, *41*(3), 400-417.

Hubbard, P. (1991). Evaluating computer games for language learning. *Simulation & Gaming*, *22*(2), 220-223.

Huberty, C. J., Jordan, E. M., & Brandt, W. C. (2005). Cluster analysis in higher education research. In J. Smart (Ed.), *Higher education: Handbook of theory and research* (pp. 437-457). Springer.

Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language. *Studies in Second Language Acquisition, 21*(2), 181-193.

Hung, H. T., Yang, J. C., Hwang, G. J., Chu, H. C., & Wang, C. C. (2018). A scoping review of research on digital game-based language learning. *Computers & Education*, *126*, 89-104.

Hung, H. C., Young, S. S. C., & Lin, C. P. (2015). No student left behind: a collaborative and competitive game-based learning environment to reduce the achievement gap of EFL students in Taiwan. *Technology, Pedagogy and Education*, *24*(1), 35-49.

Hwang, G. J., Hsu, T. C., Lai, C. L., & Hsueh, C. J. (2017). Interaction of problem-based gaming and learning anxiety in language students' English listening performance and progressive behavioral patterns. *Computers & Education*, *106*, 26-42.

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160-212.

Jones, D. M. (2020). Games in the language learning classroom: Is the juice worth the squeeze. *Ludic Language Pedagogy*, *2*, 1-36.

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Vol. 1). Statistical Tools for High-Throughput Data Analysis (STHDA).

Kazemitabar, J., Amini, A., Bloniarz, A., & Talwalkar, A. S. (2017). Variable importance using decision trees. In I. Guyon (Ed.), *Advances in neural information processing systems* (pp. 426-435). Curran Associates.

Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing*, *16*(2), 142-163.

Kim, K., Schmierbach, M. G., Chung, M. Y., Fraustino, J. D., Dardis, F., & Ahern, L. (2015). Is it a sense of autonomy, control, or attachment? Exploring the effects of in-game customization on game enjoyment. *Computers in Human Behavior*, *48*, 695-705.

Kern, R. G. (1994). The role of mental translation in second language reading. *Studies in Second Language Acquisition*, 16, 441-461.

Koda, K. (1992). The effects of lower-level processing skills on FL reading performance: Implications for instruction. *The Modern Language Journal*, *76*(4), 502-512.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press.

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning,* 57, 1-44.

Krashen, S. D. (2004). *The power of reading: Insights from the research: Insights from the research*. ABC-CLIO.

Lan, Y. J., Fang, S. Y., Legault, J., & Li, P. (2015). Second language acquisition of Mandarin Chinese vocabulary: Context of learning effects. *Educational Technology Research and Development*, *63*(5), 671-690.

Lantolf, J. P. (2011). The sociocultural approach to second language acquisition: Sociocultural theory, second language acquisition, and artificial L2 development. In D. Atkinson (Ed.), *Alternative approaches to second language acquisition* (pp. 24-47). Routledge.

Lantolf, J. P., & Poehner, M. E. (2003). Dynamic assessment of L2 development. *Journal of Applied Linguistics*, *1*, 49-74.

Lee, H., Warschauer, M., & Lee, J. H. (2017). The effects of concordance-based electronic glosses on L2 vocabulary learning. *Language Learning & Technology*, *21*(2), 32-51.

Lee, H., Warschauer, M., & Lee, J. H. (2019). Advancing CALL research via data mining techniques: Unearthing hidden groups of learners in a corpus-based L2 vocabulary learning experiment. *ReCALL*, 31(2), 135–149.

Lee, J. (1987). Comprehending the Spanish subjunctive: An information processing perspective. *The Modern Language Journal,* 71, 50-57.

Lee, J., & VanPatten, B. (2003). *Making Communicative Language Teaching Happen*. (2nd Edition). Boston, MA: McGraw Hill.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition*, 26, 35-57

Liu, M., Lee, J., Kang, J., & Liu, S. (2016). What we can learn from the data: A multiple-case study examining behavior patterns by students with different characteristics in using a serious game. *Technology, Knowledge and Learning*, 21(1), 33-57.

Liu, N., & Nation, I. S. P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal*, 16, 33–42.

Loh, C. S., & Sheng, Y. (2015). Measuring the (dis-)similarity between expert and novice behaviors as serious games analytics. *Education and Information Technologies, 20*(1), 5-19.

MacMillan, F. (2016). Assessing reading. In D. Tsagari, & J. Banerjee (Eds.), *Handbook of second language assessment* (Vol. 12, pp.113-129). de Gruyter.

Marcot, B. G., Steventon, J. D., Sutherland, G. D., & McCann, R. K. (2006). Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. *Canadian Journal of Forest Research*, *36*(12), 3063-3074.

Martin, T., Petrick Smith, C., Forsgren, N., Aghababyan, A., Janisiewicz, P., & Baker, S. (2015). Learning fractions by splitting: Using learning analytics to illuminate the development of mathematical understanding. *Journal of the Learning Sciences*, 24(4), 593-637.

Mavridis, A., & Tsiatsos, T. (2017). Game-based assessment: Investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning*, *33*(2), 137-150.

McGraw, I., Yoshimoto, B., & Seneff, S. (2009). Speech-enabled card games for incidental vocabulary acquisition in a foreign language. *Speech Communication*, *51*(10), 1006-1023.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276-282.

Mendez, G., Buskirk, T. D., Lohr, S., & Haag, S. (2008). Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests. *Journal of Engineering Education*, *97*(1), 57-70.

Miller, M., & Hegelheimer, V. (2006). The SIMs meet ESL Incorporating authentic computer simulation games into the language classroom. *Interactive Technology and Smart Education, 4,* 311-328.

Ming, Y., Ruan, Q., & Gao, G. (2013). A Mandarin edutainment system integrated virtual learning environments. *Speech Communication*, *55*(1), 71-83.

Mislevy, R., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., & Haertel, G. (2003). *Design Patterns for Assessing Science Inquiry.* SRI International.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6-20.

Mislevy, R., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, structures, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90). Erlbaum.

Morton, H., Gunson, N., & Jack, M. (2012). Interactive language learning through speech-enabled virtual scenarios. *Advances in Human-Computer Interaction*, *2012*, 1-14. doi:10.1155/2012/389523

Müller, A. (2012). Research based design of a medical vocabulary videogame. *International Journal of Pedagogies and Learning*, *7*(2), 122-134.

Patton, M. Q. (2014). *Qualitative research & evaluation methods: Integrating theory and practice*. Sage.

Palaiogiannis, A. (2014). Using video games to foster strategy development and learner autonomy within a secondary school context. *Research Papers in Language Teaching and Learning*, *5*(1), 259-277.

Peng, W., Song, H., Kim, J., & Day, T. (2016). The influence of task demand and social categorization diversity on performance and enjoyment in a language learning game. *Computers & Education*, *95*, 285-295.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11, 357–383.

Peterson, M. (2011). Digital gaming and second language development: Japanese learners interactions in a MMORPG. *Digital Culture & Educ*ation, *3*(1), 56-73.

Peterson, M. (2012). EFL learner collaborative interaction in Second Life. *ReCALL*, 24(1), 20–39.

Poole, F. J., & Clarke-Midura, J. (2020). A systematic review of digital games in second language learning studies. *International Journal of Game-Based Learning*, *10*(3), 1-15.

Poole, F., Franco, J., & Clarke-Midura, J. (2018). Developing a personalized, educational gaming experience for young Chinese DLI learners: A design-based approach. In R. Zheng (Ed.), *Digital technologies and instructional design for personalized learning* (pp. 253-274). IGI Global

Poole, F., Clarke-Midura, J., Sun, C., & Lam, K. (2019). Exploring the pedagogical affordances of a collaborative board game in a dual language immersion classroom. *Foreign Language Annals*, *52*(4), 753-775.

Poole, F., & Sung, K. (2016). A preliminary study on the effects of an E-gloss tool on incidental vocabulary learning when reading Chinese as a foreign language. *Journal of Chinese Language Teachers Association*. 51(3), 266–285.

Piirainen-Marsh, A., & Tainio, L. (2009). Collaborative game-play as a site for participation and situated learning of a second language. *Scandinavian Journal of Educational Research*, *53*(2), 167-183.

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language,* 18(1), 1-28.

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon*, *9*(5), 1-6.

Quellmalz, E. S., Davenport, J. L., Timms, M. J., DeBoer, G. E., Jordan, K. A., Huang, C. W., & Buckley, B. C. (2013). Next-generation environments for assessing and promoting complex science learning. *Journal of Educational Psychology*, 105, 1100-1114.

R Core Team (2017) *R: A language and environment for statistical computing*. https://www.R-project.org/

Rama, P. S., Black, R. W., van Es, E., & Warschauer, M. (2012). Affordances for second language learning in World of Warcraft. *ReCALL*, *24*(3), 322-338.

Ranalli, J. (2008). Learning English with The Sims: exploiting authentic computer simulation games for L2 learning. *Computer Assisted Language Learning*, *21*(5), 441-455.

Rankin, Y. A., Gold, R., & Gooch, B. (2006). 3D role-playing games as language learning tools. In *Conference proceedings of Eurographics* 2006 (pp. 33-38). ACM.

Reinders, H. (2018). Learning analytics for language learning and teaching. *JALT CALL Journal*, *14*(1), 77-86.

Reinders, H., & Wattana, S. (2014). Can I say something? The effects of digital game play on willingness to communicate. *Language Learning & Technology*. 18(2), 101-123.

Reinhardt, J. (2019). *Gameful second and foreign language teaching and learning: Theory, research, and practice*. Springer.

Reinhardt, J., & Sykes, J. (2014). Special issue commentary: Digital game and play activity in L2 teaching and learning. *Language Learning & Technology*, *18*(2), 2-8.

Roever, C. (2001). Web-based language testing. *Language Learning & Technology*, *5*(2), 84-94.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

Saldaña, J. (2015). *The coding manual for qualitative researchers*. Sage.

Saito, Y., Garza, T. J., & Horwitz, E. K. (1999). Foreign language reading anxiety. *The Modern Language Journal*, 83(2), 202-218.

Scholz, K. (2017). Encouraging free play: Extramural digital game-based language learning as a complex adaptive system. *Calico Journal*, *34*(1), 39-57.

Shen, H. H., & Jiang, X. (2013). Character reading fluency, word segmentation accuracy, and reading comprehension in L2 Chinese. *Reading in a Foreign Language*, *25*(1), 1-25. http://dx.doi.org/10.1111/j.1540-4781.2007.00511.x

Shen, H. H., & Ke, C. (2007). Radical awareness and word acquisition among nonnative learners of Chinese. *The Modern Language Journal*, *91*(1), 97-111.

Shen, H., & Tsai, C. H. (2010). A web-based extensive reading program and its assessment system. *Journal of the Chinese Language Teachers Association* 45(2), 19-47.

Shrum, J. L., & Glisan, E. W. (2010). *Teacher's handbook: Contextualized language instruction* (4th ed.). Heinle & Heinle.

Shu, H., & Anderson, R. C. (1999). Learning to read Chinese: The development of metalinguistic awareness. In J. Wang, A. W. Inhoff, & H. Chen (Eds.), *Reading Chinese script: A cognitive analysis* (pp. 1-18). Erlbaum.

Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In J. M. Spector, M. D. Merrill, J. Elen & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 311-321). Springer.

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games.* Cambridge, MA: The MIT press.

Shute, V. J., & Wang, L. (2015). Measuring problem solving skills in Portal 2. In P. Isaias, J. M. Spector, D. Ifenthaler, & D. G. Sampson (Eds.), *E-learning systems, environments and approaches* (pp. 11-24). New York: Springer.

Shute, V., Ke, F., & Wang, L. (2017). Assessment and adaptation in games. In P. Wouters & H. van Oostendorp (Ed.), *Instructional techniques to facilitate learning and motivation of serious games* (pp. 59-78). Springer.

Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254), ACM.

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130-135.

Sørensen, B. H., & Meyer, B. (2007). Serious games in language learning and teaching--A theoretical perspective. In *Proceedings of the 3rd International Conference of Digital Games Research Association* (pp. 559-566). Digital Games Research Association.

Stratman, J., F., & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols. In P. Samagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 89-112). Sage

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research techniques*. Sage.

Suh, S., Kim, S. W., & Kim, N. J. (2010). Effectiveness of MMORPG-based instruction in elementary English education in Korea. *Journal of Computer Assisted Learning*, *26*(5), 370-378.

Tan, L. H., Spinks, J. A., Eden, G. F., Perfetti, C. A., & Siok, W. T. (2005). Reading depends on writing, in Chinese. *Proceedings of the National Academy of Sciences of the United States of America*, 102(24), 8781-8785.

Therneau, T. M., & Atkinson, E. J. (1997). *An introduction to recursive partitioning using the RPART routines* [Technical report]. Mayo Foundation.

Therneau, T., & Atkinson, B. (2019*). rpart: Recursive portioning and regression tree. R package version 4.1-15*. https://cran.r-project.org/package=rpart

Thorne, S. L., Fischer, I., & Lu, X. (2012). The semiotic ecology and linguistic complexity of an online game world. *ReCALL*. 3, 279-301.

Tsui, A. B., & Fullilove, J. (1998). Bottom-up or top-down processing as a discriminator of L2 listening performance. *Applied Linguistics*, *19*(4), 432-451.

U.S. Department of State. (2020, March 6). *Foreign language training*. https://www.state.gov/foreign-language-training/

Vandercruysse, S., Vandewaetere, M., Cornillie, F., & Clarebout, G. (2013). Competition and students' perceptions in a game-based language learning environment. *Educational Technology Research and Development*, *61*(6), 927-950.

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research*, *34*(3), 229-243.

Vosburg, D. (2017). The Effects of Group Dynamics on Language Learning and Use in an MMOG. *CALICO Journal*, *34*(1), 58-74.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, J. (2009). Electronic strategies to improve Chinese reading skills. In R. D. C. V. Marriott & P. L. Torres (Eds). *Handbook of research on e-learning methodologies for language acquisition* (pp. 237-252). Information Science Reference.

Wang, J. (2012). The use of e-gloss tool to read e-text by intermediate and advanced learners of Chinese. *Computer Assisted Language Learning,* 25(5), 475-487.

Wang, H., Chang, B. R., Li, Y. S., Lin, L. H., Liu, J., & Sun, Y. L. (1986). *Xiandai Hanyu Pinlu Cidian* [Dictionary of the frequency of vocabulary in modern Chinese], Beijing Language Institute Press.

Wang, J., & Upton, T. (2012). The impact of using a pop-up dictionary on the reading process of beginning learners of Chinese. *Journal of Chinese Language Teachers Association,* 41(1), 23-41.

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. *Language Teaching*, *31*(2), 57-71.

Warschauer, M., Yim, S., Lee, H., & Zheng, B. (2019). Recent contributions of data mining to language learning research. *Annual Review of Applied Linguistics*, *39*, 93-112.

Watzinger-Tharp, J., Rubio, F., & Tharp, D. S. (2018). Linguistic performance of dual language immersion students. *Foreign Language Annals*, *51*(3), 575–595

Wiggins, G. (1998). *Educative assessment. Designing assessments to inform and improve student performance*. Jossey-Bass.

Williams, G. (2011). *Data mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer Science & Business Media.

Wickham, H. (2017). Tidyverse: Easily install and load the 'tidyverse'. *R package version*, *1*(1).

Wolf, D. F. (1993). Issues in reading comprehension assessment: Implications for the development of research instruments and classroom tests. *Foreign Language Annals*, *26*(3), 322-331.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving—More than reasoning? *Intelligence*, *40*(1), 1-14.

Xie, T., & Yao, T. (2009). Technology in Chinese language teaching and learning. In M. Everson & Y. Xiao (Eds.), *Teaching Chinese as a foreign language* (pp. 151-172). Cheng & Tsui.

Yim, S., Wang, D., Olson, J., Vu, V., & Warschauer, M. (2017, February). Synchronous collaborative writing in the classroom: Undergraduates' collaboration practices and their impact on writing style, quality, and quantity. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 468-479), ACM.

Yoshida, M. (2008). Think-aloud protocols and type of reading task: The issue of reactivity in L2 reading research. In *Selected proceedings of the 2007 second language research forum* (pp. 199-209). Somerville, MA: Cascadilla Proceedings Project.

Yorio, C. A. (1971). Some sources of reading problems for foreign-language learners. *Language Learning*, *21*(1), 107-115.

Yudintseva, A. (2015). Synthesis of research on video games for the four second language skills and vocabulary practice. *Open Journal of Social Sciences*, *3*, 81-98.

Zhao, A., Guo, Y., & Dynia, J. (2013). Foreign language reading anxiety: Chinese as a foreign language in the United States. *The Modern Language Journal*, *97*(3), 764-778. https://doi.org/10.1111/j.1540-4781.2013.12032.x

Zheng, D., Bischoff, M., & Gilliland, B. (2015). Vocabulary learning in massively multiplayer online games: context and action before words. *Educational Technology Research and Development*, *63*(5), 771-790.

Zheng, D., Newgarden, K., & Young, M. F. (2012) Multimodal analysis of language learning in a World of Warcraft play: Languaging as values-realizing. *ReCALL, 24*(3), 339-360.

APPENDICES

Appendix A

Background and Affect Surveys

PRE Survey

1. First Name	
2. Last Name	
3. Age	
4. Class (Circle one)	Red | Blue
5. Gender (Circle one)	Male | Female | Other

6. What language do you speak at home?
    a. English
    b. Chinese
    c. Spanish
    d. Other. If other, what language do you speak with your family _____.

7. Do you play video games at home?
    a. Yes
    b. No

8. How often do you play video games?
    a. Every day
    b. At least once a week
    c. At least once a month
    d. Less than once a month
    e. I have only played a few times in my life

9. How well do you the following things in Chinese?

Check one for each skill.

	0 - Not at all	1	2	3	4	5	6	7	8 – I am the best in class
Speak									
Listen									
Write									
Read									

10. Answer the following questions by checking one box for each item.

1 = I completely disagree. ----- 8 = I completely Agree

Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
It bothers me to encounter words I can't pronounce while reading Chinese								
I am worried about all the new symbols I have to learn in order to read Chinese								
I enjoy reading Chinese								
I feel confident when I am reading in Chinese								
The hardest part of learning Chinese is learning to read.								
I would be happy to learn to speak Chinese rather than having to learn to read as well.								
I get upset when I'm not sure whether I understand what I am reading in Chinese.								

Chinese Character	I know the answer	I think I know the answer	I'm just guessing	Pinyin	English
蚂蚁					
蝙蝠					
欢迎					
龙					
鹰					
摘					
熊					
布					
找					
木头					
野猪					
狼					
宝剑					
棉花					
召唤					
蛇					
告诉					
砂石					
害怕					
折					
香蕉					
蝎子					
战斗					
兵马俑					
买菜					
石头					

Chinese Character	I know the answer	I think I know the answer	I'm just guessing	Pinyin	English
无聊					
地震仪					
上海					
需要					
磁铁					
鸵鸟					
兔子					
北京					
船					
蜜蜂					
听说					
狐狸					
帮忙					
甲虫					
武器					
指南针					
知道					
喜欢					
猫头鹰					

1. 蝙蝠喜欢吃水果。 ★蝙蝠会吃哪个？ A – 猪肉 B – 苹果 C – 米饭	6. 兵马俑很古老，所以人们需要保护他们。 ★为什么兵马俑很古老？ A – 兵马俑太胖了 B – 因为生病了 C – 是中国古代的人
2. 这个指南针坏了,所以指向西边。 ★指南针应该指哪边？ A – 东边 B – 北边 C – 南边	7. 蜜蜂不能和大的动物战斗。 ★蜜蜂会和哪个动物战斗？ A – 老鼠 B – 熊 C – 狼
3. 虽然他喜欢红色的花但是他摘了黄色的花。 ★他现在有什么颜色的花？ A – 蓝色 B – 黄色 C – 红色	8.磁铁可以吸引铁的东西。 ★磁铁可以吸引哪个东西？ A – 铅笔 B – 书 C – 白板
4. 今天的天气很舒服。 ★今天的天气怎么样？ A – 不是很冷不是很热 B – 太热 C – 又刮风又下大雨	9.李老师需要你帮忙。 ★李老师可能要你做什么？ A – 收集学生的石头 B – 和朋友说话 C – 在教室睡觉

5. 其实人们不应该害怕蝙蝠，应该害怕野猪。 ★人们应该害怕哪个动物？ A – 蝙蝠 B – 狗 C – 野猪	10. 我听说中国的龙可以召唤雷电和暴雨， 而西方的龙才是一个巨大的怪物。 ★西方的龙。。。 A – 很大 B – 可以召唤雷电 C – 代表幸运

Post Survey

11. First Name	
12. Last Name	

13. How well do you do the following things in Chinese?
Check one for each skill.

	0 - Not at all	1	2	3	4	5	6	7	8 – I am the best in class
Speak									
Listen									
Write									
Read									

14. Answer the following questions by checking one box for each item.

1 = I completely disagree. ----- 8 = I completely Agree

Item	1	2	3	4	5	6	7	8
I enjoy reading Chinese								
I would be happy to learn to speak Chinese rather than having to learn to read as well.								
I felt irritated while playing the game.								
I felt bored playing the game.								
I get upset when I'm not sure whether I understand what I am reading in Chinese.								
The story of the game interested me.								
I put in a lot of effort while playing the game.								
I felt happy playing the game.								
I am worried about all the new symbols I have to learn in order to read Chinese								
The game was challenging								
It bothers me to encounter words I can't pronounce while reading Chinese								
I forgot everything around me while playing the game.								
The hardest part of learning Chinese is learning to read.								
I feel confident when I am reading in Chinese								
The game was boring.								
I felt frustrated while playing the game.								
I was totally absorbed in the game.								

Chinese Character	I know the answer	I think I know the answer	I'm just guessing	Pinyin	English
熊					
船					
战斗					
兔子					
蝙蝠					
欢迎					
木头					
喜欢					
蛇					
野猪					
上海					
知道					
龙					
宝剑					
买菜					
猫头鹰					
砂石					
蝎子					
找					
北京					
蚂蚁					
香蕉					
石头					
告诉					
地震仪					
甲虫					

Chinese Character	I know the answer	I think I know the answer	I'm just guessing	Pinyin	English
害怕					
折					
指南针					
无聊					
武器					
狐狸					
召唤					
鹰					
摘					
狼					
帮忙					
兵马俑					
鸵鸟					
棉花					
听说					
磁铁					
布					
蜜蜂					
需要					

1. 虽然他喜欢红色的花但是他摘了蓝色的花。 ★他现在有什么颜色的花？ A – 蓝色 B – 黄色 C – 红色	6. 磁铁可以吸引铁的东西。 ★磁铁不可以吸引哪个东西？ A – 钥匙 B – 书 C – 白板
2. 兵马俑很古老，所以人们需要保护他们。 ★为什么兵马俑很古老？ A – 兵马俑太胖了 B – 是中国古代的人 C –因为生病了	7. 蜜蜂只能和小的动物战斗。 ★蜜蜂会和哪个动物战斗？ A – 野猪 B – 老鼠 C – 狼
3. 今天的天气很不舒服。 ★今天的天气怎么样？ A – 不是很冷不是很热 B – 天气很不错! C –太热!	8. 我听说中国的龙可以召唤暴雨， 而西方的龙才可以喷火。 ★中国的龙…？ A – 可以喷火 B – 可以让下雨 C – 是不好的
4. 这个指南针是指向西边。 ★指南针应该指哪边？ A –北边 B – 东边 C – 西边	9. 狼喜欢吃蔬菜。 ★狼会吃哪个？ A – 胡萝卜 B – 苹果 C – 米饭

5. 其实人们不应该害怕猫头鹰，应该害怕狼。 ★人们应该害怕哪个动物？ A – 狼 B – 猫头鹰 C – 野猪	10. 李老师需要你帮忙。 ★李老师可能要你做什么？ A – 现在回家 B – 和朋友说话 C – 给校长一张纸

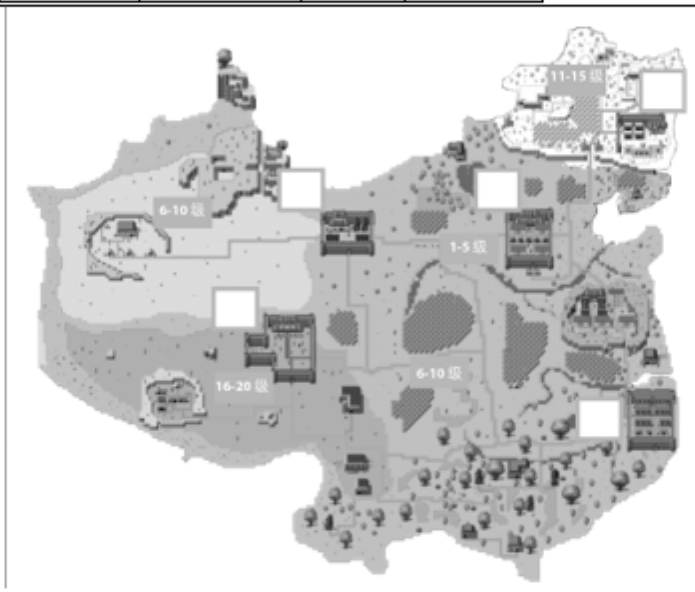
Appendix B

Workbook

龙的传说

名字_____

北京	成都	上海	西安	哈尔滨
Běijīng	Chéngdū	Shànghǎi	Xī'ān	Hāěrbīn



你觉得你应该先去哪儿? 然后呢? 为什么?

汉子	拼音	写		
北京	Běijīng			
成都	Chéngdū			
上海	Shànghǎi			
西安	Xī'ān			
哈尔滨	Hāěrbīn			

第二天

汉字	拼音	写		
物品	wùpǐn			
技能	jìnéng			
装备	zhuāngbèi			
选择	xuǎnzé			
保存	bǎocún			
游戏结束	yóuxìjiéshù			

物品	装备
物品= wùpǐn	武器=wǔqì
战斗卡= zhàndòu kǎ	盾牌= dùnpái
物品卡= wùpǐnkǎ	头上的= tóushàngde
	身体的= shēntǐde

在哪儿可以找到这些东西?

			
zhǐnánzhēn	bǎojiàn	xiāngjiāo	Mǎyǐ kǎ
			
dì tú	dùnpái	shàngyī	húluóbo
			
Bù kǎ	Yīng kǎ	Shāshí kǎ	màozi

汉字	指南针	蚂蚁卡	胡萝卜	上衣	宝剑	布卡
拼音						
汉子	盾牌	地图	香蕉	砂石卡	鹰卡	帽子
拼音						

第三天

装备: zhuāngbèi

汉字	拼音	写		
宝剑	bǎojiàn			
盾牌	dùnpái			
帽子	màozi			
上衣	shàngyī			
截止	jiézhǐ			
鞋子	xiézi			

你觉得上面装备会提高你的攻击,防守,还是敏捷度?

攻击	防守	敏捷度
gōngjī	fángshǒu	Mǐnjié dù
		

在游戏里找下面的装备, 然后记一下它怎么帮助你

装备: zhuāngbèi	颜色	攻击	防守	敏捷度
宝剑				
盾牌				
帽子				
上衣				
截止				
鞋子				

1. 装备有哪些颜色?
2. 哪个颜色最厉害?
3. 你在哪儿可以找到装备?

第四天

你觉得哪些动物厉害?	你觉得它们有什么特能力?
你害怕哪些动物?	你觉得它们喜欢吃什么?

	等级范围: 1-5 喜欢吃: 糖果 特能力: 咬		等级范围: 喜欢吃: 特能力:
	等级范围: 喜欢吃: 特能力:		等级范围: 喜欢吃: 特能力:
	等级范围: 喜欢吃: 特能力:		等级范围: 喜欢吃: 特能力:
	等级范围: 喜欢吃: 特能力:		等级范围: 喜欢吃: 特能力:
	等级范围: 喜欢吃: 特能力:		等级范围: 喜欢吃: 特能力:

第五天

你觉得哪些动物厉害?	你觉得它们有什么特能力?
你害怕哪些动物?	你觉得它们喜欢吃什么?

	等级范围: 喜欢吃: 特能力:		等级范围: 喜欢吃: 特能力:
	等级范围: 喜欢吃: 特能力:		等级范围: 喜欢吃: 特能力:
	等级范围: 喜欢吃: 特能力:		等级范围: 喜欢吃: 特能力:
	等级范围: 喜欢吃: 特能力:		等级范围: 喜欢吃: 特能力:
	等级范围: 喜欢吃: 特能力:		

第六天

汉字	拼音	写		
木头	mùtóu			
石头	shítóu			
砂石	shāshí			
线	xiàn			
刺	cì			
胡萝卜	Húluóbo			
香蕉	xiāngjiāo			
西红柿	xīhóngshì			
葡萄	pútáo			
苹果	píngguǒ			

和你的同学说下面的中文单词

东西	卖多少钱?
木头	
石头	
砂石	
线	
刺	
胡萝卜	
香蕉	
西红柿	
葡萄	
苹果	

1. 你觉得你怎么能用这些东西?
2. 你在哪儿能找到这些东西?
3. 哪些东西最贵? 最便宜?

第七天

药水	拼音		多少钱?
防御药水	Fángyù yàoshuǐ		
生命药水	Shēngmìng yàoshuǐ		
增强药水	Zēngqiáng yàoshuǐ		
解毒药水	Jiědú yàoshuǐ		
加速药水	Jiāsù yàoshuǐ		
魔法药水	Mófǎ yàoshuǐ		

1. 你什么时候会用上面的药水?
2. 它们有什么用?
3. 你在哪儿能买?

重要的人物

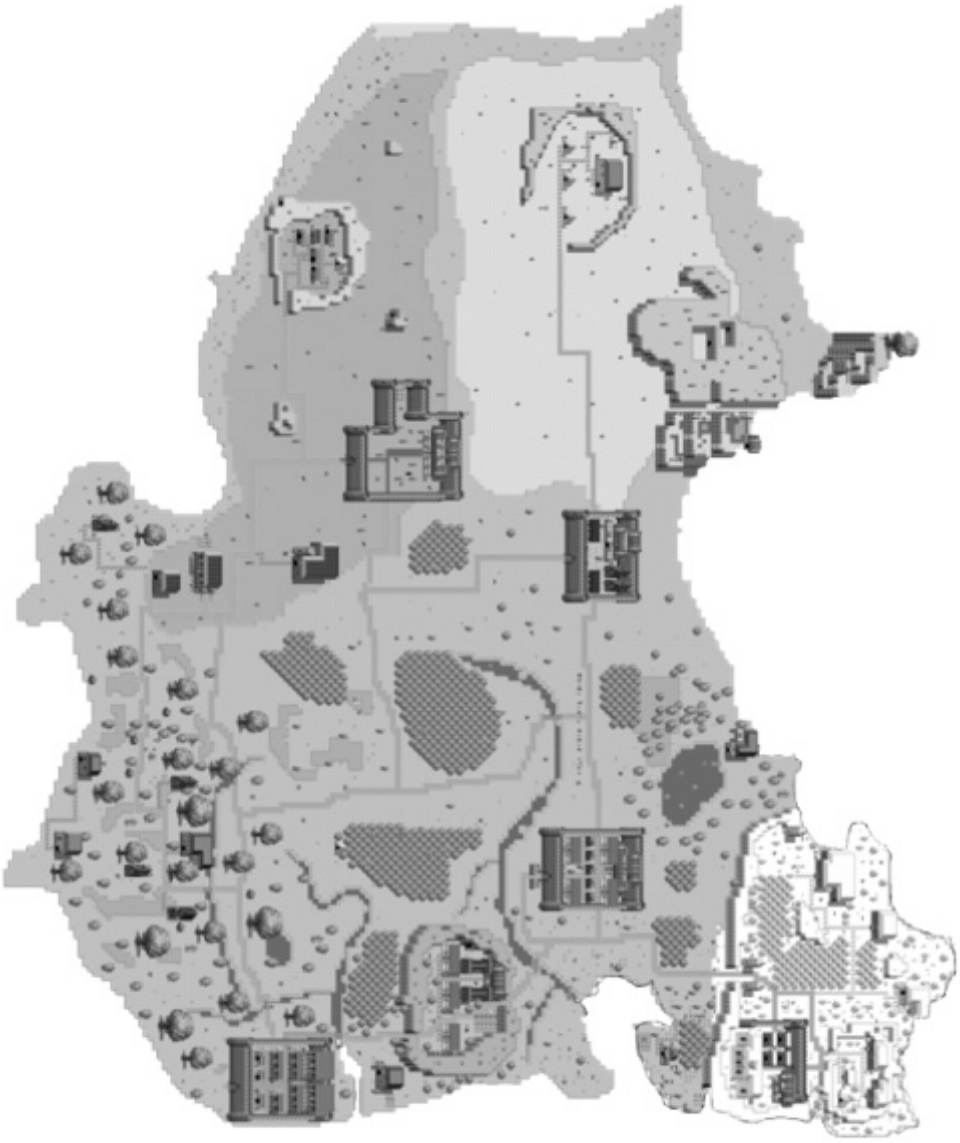
司马迁- sīmǎqiān	
在哪儿--	做什么 ---
兵马俑- bīngmǎyǒng	
在哪儿--	做什么 ---
武则天- wǔzétiān	
在哪儿--	做什么 ---
孔子- kǒng zǐ	
在哪儿--	做什么 ---
孙子- sūn zǐ	
在哪儿--	做什么 ---
郑和- zhènghé	
在哪儿--	做什么 ---
张衡- zhānghéng	
在哪儿--	做什么 ---

第八天

任务一 谁: 在哪儿: 做什么:	任务七 谁: 在哪儿: 做什么:
任务二 谁: 在哪儿: 做什么:	任务八 谁: 在哪儿: 做什么:
任务三 谁: 在哪儿: 做什么:	任务九 谁: 在哪儿: 做什么:
任务四 谁: 在哪儿: 做什么:	任务十 谁: 在哪儿: 做什么:
任务五 谁: 在哪儿: 做什么:	任务十一 谁: 在哪儿: 做什么:
任务六 谁: 在哪儿: 做什么:	任务十二 谁: 在哪儿: 做什么:

其他的人物

名字: 在哪儿: 说什么: 描述:	名字: 在哪儿: 说什么: 描述:
名字: 在哪儿: 说什么: 描述:	名字: 在哪儿: 说什么: 描述:
名字: 在哪儿: 说什么: 描述:	名字: 在哪儿: 说什么: 描述:
名字: 在哪儿: 说什么: 描述:	名字: 在哪儿: 说什么: 描述:
名字: 在哪儿: 说什么: 描述:	名字: 在哪儿: 说什么: 描述:
名字: 在哪儿: 说什么: 描述:	名字: 在哪儿: 说什么: 描述:
名字: 在哪儿: 说什么: 描述:	名字: 在哪儿: 说什么: 描述:
名字: 在哪儿: 说什么: 描述:	名字: 在哪儿: 说什么: 描述:



Appendix C

Informed Consent

OFFICE of RESEARCH and GRADUATE STUDIES
UtahStateUniversity
INSTITUTIONAL REVIEW BOARD

Page 1 of 3
Protocol #9373
IRB Approval Date: 05/18/2018
Consent Document Expires: 05/17/2019
IRB Password Protected per IRB Coordinator

v.8.3: May2017

Informed Consent

INTRODUCTION

Your child is asked to participate in a research study conducted by Frederick Poole and Dr. Jody Clarke-Midura from the Instructional Technology and Learning Sciences Program at Utah State University (USU). Your child was selected as a possible participant in this study because they are in the 5th grade Dual Language Immersion (DLI) Program at Cedar Ridge Elementary. We are interested in investigating how a digital game can be used as a means of assessing your child's language skills while they play a game. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not your child should participate.

Procedures

Your child's participation in this study is completely voluntary and he/she is free to choose whether to participate or not. The choice to participate or not will have no impact on your child's grades. We have invited students in both 5th grade Chinese classes (about 40 students) to participate. Those who participate will play the game during 7 class periods for approximately 6.5 hours. Before and after the study, students will complete a 20-minute language assessment. This study will be completed in 4 weeks.

If your child volunteers to participate in this study, we would ask him/her to do one or more of the following things:

- Complete vocabulary and reading comprehension assessments.
- Play an educational video game during computer time at school.

Voluntary Participation and Withdrawal

If you provide consent for your child to participate in this study and your child also assents to be in this study, he/she may subsequently withdraw from it at any time without penalty or consequences of any kind. If a child withdraws from the study, he/she will still be able to play the game, but data will not be collected. While playing the game we will collect data that captures how long your child reads in-game texts, the type of assistance that your child uses while playing the game, the number of times that your child interacts with in-game objects while playing. If your child withdraws from this study during data collection we will delete all data collected, however if withdrawal occurs after the data has been de-identified we will not be able to destroy the data collected.

Risks

This is a minimal risk research study. That means that the risks of participating are no more likely or serious than those you encounter in everyday activities. However, some students may experience frustration while playing the game or be uncomfortable being audio-recorded. In order to minimize those risks and discomforts, the research team will check-in with the teacher periodically to determine if any students are uncomfortable playing the game or are noticeably uncomfortable being audio-recorded. If the learner experiences discomfort and/or frustration as a result of playing the game, the research assistant and/or instructor will first attempt to calm the child, if this does not work, the child will be removed from the game to reduce discomfort and/or frustration. In addition, there is a small risk of loss of confidentiality. If you have a bad research-related experience or are injured in any way during your participation, please contact the principal investigator of this study right away at 435-797-0571 or jody.clarke@usu.edu.

Benefits

We believe that your child may develop reading skills and Chinese vocabulary by playing the game, which could help them with their studies. In addition, we believe that your child will enjoy playing the game.

Furthermore, results of this study will help promote better teaching methodology and assessments in the dual language immersion program. Currently, there is little research regarding best practices for simultaneously teaching and assessing language skills one activity. This research will examine how a digital game can assess learners while they learn and play. However, we cannot promise any direct benefits associated with participation in this study.

Confidentiality

Research records will be kept confidential, consistent with federal and state regulations. Only the investigator and student investigator will have access to the data which will be kept in a locked file cabinet or on a password protected computer in a locked room. Students will use pseudonyms to log into the game and thus their data will remain anonymous. Audio recorded files will be coded with a pseudonym and/or ID number as soon as they are transcribed. A digital word document that links real names to pseudonyms will be kept in a restricted access folder on Box.com. Once the data has been transcribed, the identifying word document will be destroyed, which will occur no later than May 2020. To protect your child's privacy, personal, identifiable information will be removed from study documents and replaced with a study identifier. Identifying information will be stored separately from data and will be kept until all data is cleaned and assigned an ID number. Then all identifiable information will be destroyed by May 1st 2020. These de-identified data may be used or distributed for future research without additional consent from you. If you do not wish for us to use your child's information in this way, please state so below.

Voluntary Participation

Your child's participation in this research is completely voluntary. However, if you agree to allow your child to participate now and change your mind later, and if we have already collected information, we will not be able to remove your child's data since the data is anonymous. If you decide not to allow your child to participate, the services you receive from researcher and school will not be affected in any way. The researchers may choose to terminate your child's participation in this research study if they do not follow the directions on the handout.

IRB Review

The Institutional Review Board (IRB) for the protection of human research participants at Utah State University has reviewed and approved this study. If you have questions about the research study itself, please contact the Principal Investigator at (435) 797-0571 or jody.clarke@usu.edu. If you have questions about your rights or would simply like to speak with someone other than the research team about questions or concerns, please contact the IRB Director at (435) 797-0567 or irb@usu.edu.

Please replace this line with an electronic signature

Jody Clarke-Midura
Principal Investigator
435-797-0571; jody.clarke@usu.edu

Please replace this line with an electronic signature

Frederick J Poole
Student Investigator
719-232-8659; frdbrick@gmail.com

Informed Consent

By signing below, you agree to allow your child to participate in this study. You indicate that you understand the risks and benefits of participation, and that you know what you will be asked of your child to do. You also agree that you have asked any questions you might have, and are clear on how to stop your participation in the study for your child if you choose to do so. Please be sure to retain a copy of this form for your records.

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to allow my child to participate in this study. I have been given a copy of this form.

I consent to allow my child to participate in the following activities (please check boxes next to activities you will allow):

- ☐ Complete vocabulary and reading comprehension assessments
- ☐ Play an educational video game during computer time at school.
- ☐ Be audio-recorded while playing the game.

_____
Name of Participant (Student)

_____
Name of Legal Representative (Parent/Guardian)

_____ _____
Signature of Legal Representative (Parent/Guardian) Date

☐ I do **not** agree to allow my de-identified information/biospecimens to be used or shared for future research.

Youth Assent

We are doing a research study about playing digital games for learning Chinese. Research studies help us learn more about people. If you would like to be a part of this research study, you will play an educational video game in class. We will also ask you some questions about Chinese.

Before you agree to do these things, we need to tell you a little more. First, when you play the game you may become frustrated. If you are in this study, there are also some things that you may like, such as playing the game or learning new words. Also, we will tell other people about what we learned from doing this study with you and the 40 other people who are in the study, but we won't tell anyone your name or that you were in the study.

If this sounds like something you would like to do, we will ask you to say that you understand what we talked about, and that you do want to participate. You do not have to be in this study if you do not want to be. If you decide to stop after we begin just tell Fred or your teacher and you don't have to continue, that's okay, too. No one will be upset if you don't want to do this, or change your mind later.

You can ask any questions you have, now or later. Your parents know about this research study, and they have said you can participate, if you want.

If you would like to be in this study, please sign your name and write the date.

_____ _____
Name Date

Appendix D

R Libraries Used for Data Cleaning and Analysis

Library	Citation	Application
tidyverse	Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686	These applications were used for data cleaning/ manipulation.
plyr	Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.	
expss	Gregory Demin (2020). expss: Tables, Labels and Some Useful Functions from Spreadsheets and 'SPSS' Statistics. R package version 0.10.2. https://CRAN.R-project.org/package=expss	
chron	David James and Kurt Hornik (2020). chron: Chronological Objects which Can Handle Dates and Times. R package version 2.3-55.	
lubridate	Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.	
RcppRoll	Kevin Ushey (2018). RcppRoll: Efficient Rolling / Windowed Operations. R package version 0.3.0. https://CRAN.R-project.org/package=RcppRoll	
dplyr	Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr	
tidyr	Hadley Wickham and Lionel Henry (2020). tidyr: Tidy Messy Data. R package version 1.0.2. https://CRAN.R-project.org/package=tidyr	
data.table	Matt Dowle and Arun Srinivasan (2019). data.table: Extension of `data.frame`. R package version 1.12.8. https://CRAN.R-project.org/package=data.table	
questionr	Julien Barnier, François Briatte and Joseph Larmarange (2018). questionr: Functions to Make Surveys Processing Easier. R package version 0.7.0. https://CRAN.R-project.org/package=questionr	
tibble	Kirill Müller and Hadley Wickham (2020). tibble: Simple Data Frames. R package version 3.0.0. https://CRAN.R-project.org/package=tibble	
textclean	Rinker, T. W. (2018). textclean: Text Cleaning Tools version 0.9.3. Buffalo, New York. https://github.com/trinker/textclean	These libraries were primarily used for cleaning text data specifically.
tm	Ingo Feinerer and Kurt Hornik (2019). tm: Text Mining Package. R package version 0.7-7. https://CRAN.R-project.org/package=tm	
stringr	Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr	

Library	Citation	Application
tmcn	Jian Li (2019). tmcn: A Text Mining Toolkit for Chinese. R package version 0.2-13. https://CRAN.R-project.org/package=tmcn	
tidytext	Silge J, Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." _JOSS_, *1*(3). doi: 10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.	
topicmodels	Grün B, Hornik K (2011). "topicmodels: An R Package for Fitting Topic Models." _Journal of Statistical Software_, *40*(13), 1-30. doi: 10.18637/jss.v040.i13 (URL: https://doi.org/10.18637/jss.v040.i13).	
readr	Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr	
jiebaR	Qin Wenfeng and Wu Yanyi (2019). jiebaR: Chinese Text Segmentation. R package version 0.11. https://CRAN.R-project.org/package=jiebaR	
quanteda.corpora	Kenneth Benoit (2020). quanteda.corpora: A collection of corpora for quanteda. R package version 0.91. http://github.com/quanteda/quanteda.corpora	
quanteda	Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). "quanteda: An R package for the quantitative analysis of textual data." _Journal of Open Source Software_, *3*(30), 774. doi: 10.21105/joss.00774 (URL: https://doi.org/10.21105/joss.00774), <URL: https://quanteda.io>.	
rpart	Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart	These were all used for the CART analyses and visualization of the decision trees in Chapter 3.
rattle	Williams, G. J. (2011), Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Use R!, Springer.	
rpart.plot	Stephen Milborrow (2019). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.0.8. https://CRAN.R-project.org/package=rpart.plot	
rcolorBrewer	Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer	
cluster	Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.	These packages was used to run the cluster analysis in Chapter 3.
factoextra	Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra	

Library	Citation	Application
ggplot2	H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.	These packages were used for most of the visualizations in this dissertation.
extrafont	Winston Chang, (2014). extrafont: Tools for using fonts. R package version 0.17. https://CRAN.R-project.org/package=extrafont	
shades	Jon Clayden (2019). shades: Simple Colour Manipulation. R package version 1.4.0. https://CRAN.R-project.org/package=shades	
plotly	C. Sievert. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida, 2020.	
hrbrthemes	Bob Rudis (2020). hrbrthemes: Additional Themes, Theme Components and Utilities for 'ggplot2'. R package version 0.8.0. https://CRAN.R-project.org/package=hrbrthemes	
rstatix	Alboukadel Kassambara (2020). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.4.0. https://CRAN.R-project.org/package=rstatix	These packages along with base R were used for most of the analysis and descriptive statistics, including: correlations, effect size, and Cronbach alpha.
psych	Revelle, W. (2019) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 1.9.12.	
PerformanceAnalytics	Brian G. Peterson and Peter Carl (2020). PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis. R package version 2.0.4. https://CRAN.R-project.org/package=PerformanceAnalytics	
stargazer	Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer	
effsize	Torchiano M (2020). _effsize: Efficient Effect Size Computation_. doi: 10.5281/zenodo.1480624 (URL: https://doi.org/10.5281/zenodo.1480624), R package version 0.7.9, <URL: https://CRAN.R-project.org/package=effsize>.	
ggfortfiy	Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.	

Appendix E

Sample Log Data in JSON Format

“Map_id” : “2,”
{ “Pathing” : [
[“00:00:27,” Transfer, 23,11],
[“00:00:28,” Text, 你醒了！你叫什么名字？,undefined],
[“00:00:29,” Respond, \n[1],李伟,张敏,我不知道！,0],
[“00:00:29,” Text, 好名字！你是不是刚刚到了北京？,undefined],
[“00:00:29,” LookUp,名字,undefined],
[“00:00:30,” Respond, 是的！,0],
[“00:00:30,” Text, 那， 你应该去找\c[10]张伟\c！,他会教你怎么战斗(zhàndǒu)！,undefined],
[“00:00:33,” Path, 22,11],
[“00:00:33,” Path, 21,11],
[“00:00:33,” Path, 20,11],
[“00:00:33,” Path, 20,12],
[“00:00:34,” Path, 19,12],
[“00:00:34,” Path, 18,12],
[“00:00:34,” Path, 17,12],
[“00:00:34,” Path, 16,12],
[“00:00:34,” Path, 15,12],
[“00:00:34,” Path, 15,11],
[“00:00:35,” Path, 14,11],
[“00:00:35,” Path, 13,11],
[“00:00:35,” Path, 12,11],
[“00:00:35,” Path, 11,11],
[“00:00:35,” Path, 11,12],
[“00:00:35,” Path, 11,13],
[“00:00:36,” Path, 10,13],
[“00:00:36,” Path, 9,13],
[“00:00:36,” Path, 8,13],
[“00:00:37,” Path, 8,14],
[“00:00:37,” Path, 8,15],
[“00:00:37,” Path, 7,15],
[“00:00:37,” Path, 6,15],
[“00:00:37,” Path, 5,15],
[“00:00:37,” Path, 4,15],
[“00:00:38,” Event_Start, OldMan,undefined],
[“00:00:38,” Text, 你在这儿可以买书! ,undefined],
[“00:00:44,” Path, 5,15],
[“00:00:44,” Path, 5,16],
[“00:00:44,” Path, 5,17],
[“00:00:45,” Path, 5,18],
[“00:00:45,” Path, 6,18],
[“00:00:45,” Path, 7,18],
[“00:00:45,” Path, 7,19]

CURRICULUM VITAE

FREDERICK J POOLE

2830 Old Main Hill
Utah State University, Logan, UT 84322
Website: fredpoole.github.io

Phone: 719-232-8659
E-mail: frdbrick@gmail.com

EDUCATION

2015 – Present	Ph.D. Student Instructional Technology and Learning Sciences, Utah State University (Expected, May 2020) Advisor: Assistant Professor Jody Clarke-Midura, Instructional Technology and Learning Sciences Department
2015	Master of Second Language Teaching, Utah State University
2007	B.A. Double Major Spanish Literature and Psychology, University of Colorado at Boulder

PUBLICATIONS

Second Language Acquisition: Journal Articles (refereed)

2020 **Poole, F.,** & Clarke-Midura, J. A systematic review of digital games in second language learning studies. International Journal of Game Based Learning, (CiteScore = 1.27).

2019 **Poole, F.,** Clarke-Midura, J., Sun, C, & Lam, K. Exploring the pedagogical affordances of a collaborative board game in a dual language immersion classroom. *Foreign Language Annals*. 42(4), 753-775 (Impact Factor= 1.78).

2019 Walker, J., & **Poole, F.** Effects of delaying character instruction in a Chinese as second language middle school classroom. *Researching and Teaching Chinese as a Foreign Language*,2(2), 261-280.

2018 Thoms, J., Arshavskaya, E., & **Poole, F.** Open educational resources and ESL education: Insights from educators in the U.S. *TESL-EJ*, 22(2).

2018 Thoms, J., & **Poole, F.** Exploring digital literacy practices via L2 social reading, Special Issue of *L2 Journal*, *10*(2).

2017 Sung, K. Y., & **Poole, F**. Investigating the use of a smartphone social networking application on language learning. *JALT CALL Journal*, *13*(2), 97–115. (CiteScore= .17)

2017 Thoms, J., Sung, K. Y., & **Poole, F.** Investigating the linguistic and pedagogical affordances of an L2 open reading environment via *eComma*: An exploratory study in a Chinese language course. *System*, *69*, 38–53. (CiteScore = 1.55)

2017 Thoms, J., & **Poole, F**. Investigating linguistic, literary, and social affordances of L2 collaborative reading. *Language Learning & Technology*, *21*(2), 139–156. (CiteScore = 2.61)

2016 **Poole, F**., & Sung, K. A preliminary study on the effects of an E-gloss tool on incidental vocabulary learning when reading Chinese as a foreign language. *Journal of Chinese Language Teachers Association*. *51*(3), 266–285.

2016 Sung, K. Y., & **Poole, F.** Differences between native and non-native English-speaking teachers in China from the perspectives of Chinese EFL students. *The Journal of English as an International Language*, *11*(2), 1–18.

2015 Sung, K. Y., & **Poole, F.** Evaluating the impact of graded readings on the recognition of Chinese characters and reading comprehension by learners of Chinese as a foreign language. *Konin Language Studies, 3*(3), 271–294.

2015 **Poole, F.**, & Sung, K. Three approaches to beginning Chinese instruction and their effects on oral development and character recognition. *Eurasian Journal of Applied Linguistics*, *1*(1), 59–75.

<u>Second Language Acquisition: Book Chapters (refereed)</u>

Forthcoming Thoms, J. & **Poole, F.** Language teachers and the open education movement: A national survey. *Open education and foreign language learning and teaching: The rise of a new knowledge ecology.* Bristol,

UK: Multilingual Matters.

2018 **Poole, F**., Franco, J., & Clarke-Midura, J. Developing a personalized, educational gaming experience for young Chinese DLI learners: A design-based approach. In R. Zheng (Ed.), *Digital Technologies and Instructional Design for Personalized Learnin*g (pp. 253-274). IGI Global.

2016 Sung, K. Y., & **Poole, F**. Differences between native and non-native Chinese speaking teachers: Voices from overseas students who study Chinese in China. In C. P. Chou & J. Spangler (Eds.) *Chinese Education Models in a Global Age* (pp. 133–147). Springer, Singapore.

Educational Technology: Journal Articles and Conference Proceedings (refereed)

2020 Lee, V. R., **Poole, F**., Clarke-Midura, J., Recker, M., & Rasmussen, M. Introducing coding through tabletop board games and their digital instantiations across elementary classrooms and school libraries. In *Proceedings of ACM Technical Symposium on Computer Science Education (pp. xxx)*. New York: ACM.

2020 Lee, V. R., Poole, F., Rasmussen, M., Clarke-Midura, J., & Recker, M. Examining Variations in Teacher Talk when Implementing New Unplugged-to-plugged Computing Instruction. Paper presented at the *American Educational Research Association Conference*, San Francisco, CA.

2019 Clarke-Midura, J., Sun, C., Pantic, K., **Poole, F.**, & Allan, V. Using Informed Design in Informal Computer Science Programs to Increase Youths' Interest, Self-efficacy, and Perceptions of Parental Support. *ACM Transactions on Computing Education (TOCE)*, 19 (4), 1-12 (Article 37), doi: 10.1145/3319445

2018 Pantic, K., Clarke-Midura, J**., Poole, F.,** Roller, J., & Allan V. (2018). Drawing a computer scientist: Stereotypical representations or lack of awareness? *Computer Science Education Journal, 28,* 3. doi: 10.1080/08993408.2018.1533780

2018 Clarke-Midura, J., **Poole, F.**, Pantic, K., Sun, C., Allan, V. (2018). How mother and father support affects youths' interest in computer science. *Proceedings of ACM ICER conference,* Espoo, Finland, August, 2018.

2018 Clarke-Midura, J., **Poole, F.**, Pantic, K., & Allan, V. Playing mentor: A new strategy for recruiting young women into computer science. *Journal of Women and Minorities in Science and Engineering, 23*(3). (CiteScore= 1.15)

2017 Clarke-Midura, J., **Poole, F**., Pantic, K., Hamilton, M., Sun, C., & Allan, V. How Near Peer Mentoring Affects Middle School Mentees. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (SIGCSE '18). ACM, New York, NY, USA, 664-669. (Second Best Paper CS Education Research)

WORK IN PROGRESS

Second Language Acquisition: Journal Articles (refereed)

In Preparation Thoms, J. & **Poole, F.** Analyzing Digital Social Reading Practices in an L2 Chinese Literature Course.

Educational Technology: Journal Articles (refereed)

In Preparation **Poole, F.** & Lee, V. Exploring data practices by a collegiate esports team.

Under Review Clarke-Midura, J.**,** & **Poole, F.** Conceptualizing student perceptions of computer scientist from an imagined communities' framework: A keyword analysis.

Under Review Clarke-Midura, J., Pantic, K., **Poole, F.,** Allan, V., Dorward, J. Benefits of near-peer mentoring for females in computer programming: A Three-Tier Model.

PROFESSIONAL AND ACADEMIC POSITIONS

2019 -- Present Adjunct Instructor, Instructional Technology and Learning Sciences, Utah State University, Logan, UT

2017 Adjunct Instructor, School of Teacher Education and Leadership, Utah State University, Logan, UT

2015 – Present Graduate Research Assistant, Instructional Technology and Learning Sciences Department, Utah State University, Logan, UT

2013 – 2016 Graduate/Adjunct Chinese Instructor, Languages, Philosophy, and Communication Studies, Department, Utah State University, Logan, UT

2010 – 2013	Director of Studies, Mercan English School, Beijing/Taiyuan, China
2008 – 2013	English as a Foreign Language (EFL) Teacher, Mercan English School, Taiyuan, China
2007 – 2008	Assistant Dual Language Immersion Teacher (English Teacher), Government of Andalucía, Lora del Río, Seville, Spain
2006	English Teaching Assistant, English Opens Doors Program, Santiago, Chile

TEACHING

Utah State University

Spring 2020	Games and Learning (ITLS 6730) (Graduate Level Online Course)
Fall 2019	Mobile Design and Development for Learning (ITLS 3870/5750) (Graduate Level Online Course)
Spring 2019	Games and Learning (ITLS 6730) (Graduate Level Online Course)
Fall 2017	Learning Theory (ITLS 6540) (**Graduate-level course, taught last 7 weeks of course at request of department due to maternity leave of a faculty member)*
Fall 2017	Second Language/Literacy Acquisition and Development (TEAL 4745)
Fall 2013, 2014, 2015, 2016	First-Year Chinese 1 (Chinese 1010)
Spring 2015	Teaching Chinese as a Foreign Language (Chinese 4100)
Spring 2014	First-Year Chinese II (Chinese 1020

Mercan School (Private EFL Institute in China)

2008 – 2013	TOEFL/IELTS Test Prep
2008 – 2013	High School American Literature/American Culture Course
2008 – 2013	Elementary Level English Course

CONFERENCE PARTICIPATION

Presenter

2019 Using a board game in a Chinese Dual Language immersion classroom, American Council on the Teaching of Foreign Languages (ACTFL) Conference, Washington, DC.

2019 A Chinese RPG Game, Computer Assisted Language Instruction (CALICO) Conference [Showcase Presentation], Montreal Canada

2018 Analyzing digital literacy practices via L2 collaborative reading. WorldCALL, Concepción, Chile (with Dr. Joshua Thoms).

2018 Interactive stories with Twine
Computer Assisted Language Instruction Consortium (CALICO) Conference [Workshop], Urbana-Champaign, IL.

2018 Using digital games in the language classroom. Utah Foreign Language Teaching Association (UFLA) Conference, Ogden, UT.

2017 Investigating linguistic, literary, and social affordances of L2 collaborative/social reading. American Association of Applied Linguistics (AAAL) Conference, Portland, OR (with Dr. Joshua Thoms).

2017 Using a story-driven game in the Chinese dual language immersion classroom. Utah Foreign Language Teaching Association (UFLA) Conference, Orem, UT.

2016 A cross-language analysis of online language tutors' corrective feedback and learners' uptake when learning via videoconferencing tool. American Association of Applied Linguistics (AAAL) Conference, Orlando,FL.

2016 Collaboration and learning with a board game in a Chinese dual language immersion classroom. Utah Foreign Language Teaching Association (UFLA) Conference, Orem, UT.

2015 Investigating the effectiveness of 'We Chat' on language learning. American Council on the Teaching of Foreign Languages (ACTFL) Conference, San Diego, CA.

2015 The effects of E-dictionaries on incidental vocabulary learning. American Council on the Teaching of Foreign Languages (ACTFL) Conference, San Diego, CA.

2015 A new model for Chinese compliment responses. Western Conference Association for Asian Studies (WCAAS) conference, Salt Lake City, UT.

2015 A comparison and analysis of three-character instruction approaches. Chinese American Educational Research and Development Association (CAERDA) Conference, Chicago, IL.

2015 Practical applications for web-based tools in the Chinese classroom. Utah Foreign Language Teaching Association (UFLA) Conference, Ogden, UT.

2014 Graded readers in the Chinese classroom. American Council on the Teaching of Foreign Languages (ACTFL) Conference, San Antonio, TX.

2014 Effects of e-glosses on long-term retention of vocabulary when reading Chinese. Lackstrom Symposium, Logan, UT.

2013 Integrating reading into the communicative classroom. Intermountain Teachers of English to Speakers of Other Languages (I-TESOL) Conference, Salt Lake City, UT.

Contributor

2020 Lee, V. R., **Poole, F**., Clarke-Midura, J., & Recker, M. Design of an expansively-framed board game-based unit to introduce computer programming to upper elementary students. Poster presented at the *2020 International Conference of the Learning Sciences*, Nashville, TN.

PROFESSIONAL ACTIVITIES/SERVICE

2020 Manuscript Reviewer, *Ludic Language Pedagogy*

2019 Manuscript Reviewer, *Foreign Language Annals*

2019 Book Reviewer, *Language Learning & Technology*

2018 Manuscript Reviewer, *Computers & Education*

2018 Search Committee for Tenure-Track Position in Department of Instructional Technology and Learning Sciences at Utah State University.

2018 Search Committee for Professor of Practice Position in Department of Instructional Technology and Learning Sciences at Utah State University.

2017 Manuscript Reviewer, *Computers & Education*

2017 Manuscript Reviewer, *Digital Technologies and Instructional Design for Personalized Learnin*g

AWARDS & GRANTS

2019 Doctoral Student Researcher of the Year, Department of Instructional Technology and Learning Sciences, Utah State University

2018 ITLS Research & Development Scholarship (Awarded $935.00), Utah State University

2018 Doctoral Student Researcher of the Year, Department of Instructional Technology and Learning Sciences, Utah State University

2017 Graduate Enhancement Award (Awarded $4,000), Utah State University

2016 Rick Q. Lawson Scholarship (Awarded $3,000), Emma Eccles Jones College of Education and Human Services, Utah State University

2016 Presidential Fellowship (Awarded $20,000), Department of Instructional Technology and Learning Sciences, Utah State University

2015 Rick Q. Lawson Scholarship (Awarded $1,750), Emma Eccles Jones College of Education and Human Services, Utah State University

2015 Talk Abroad – Short-Term Research Grant – (Awarded $5,000 TalkAbroad credit; $1,500 for research)

2015 Pat Buckner Award for Collaborative Teaching Projects (Awarded $950), Utah Foreign Language Association

2014 Graduate Researcher of the Year, Department of Languages, Philosophy, and Communication Studies, Utah State University

2014 Graduate Research Paper of the Year, Department of Languages, Philosophy, and Communication Studies, Utah State University

2013 Graduate Researcher of the Year, Department of Languages, Philosophy, and Communication Studies, Utah State University

PROFESSIONAL AFFILIATIONS

American Educational Research Association (AERA)
Computer Assisted Language Instruction Consortium (CALICO)
American Association of Applied Linguistics (AAAL)
American Council on The Teaching of Foreign Languages (ACTFL)
Chinese Language Teachers Association (CLTA)
Teachers of English to Speakers of Other Languages Association (TESOL)
Utah Foreign Language Association (UFLA)

LANGUAGES

English – Native
Chinese (Mandarin) – Near native
Spanish – Advanced