Department of Modern Languages and Literatures and

English Studies

M.A. Thesis

It's Not Whether You Win or Lose:

Investigating the Use of Serious Games and L2 Reading Development

Author: David S. Israelsson

Supervisor: Dr. Roger Gilabert Guerrero

Academic year: 2019-20

UNIVERSITAT DE BARCELONA

Facultat de Filologia i Comunicació
Dept. Llengües i Lit. Modernes i Estudis Anglesos

Gran Via de les Corts Catalanes, 585
08007 Barcelona, SPAIN
Tel. +34 934 035 686
Fax +34 933 171 249

Màster Oficial en Lingüística Aplicada
i Adquisició de Llengües en Contextos Multilingües
LAALCM

Roger Gilabert Guerrero (nom i cognoms) com a supervisor/a del treball (Tesina de Màster) presentat com a requeriment per a l'avaluació de l'assignatura **Projecte de Recerca en Lingüística Aplicada**

presentat per l'alumne/a: David S. Israelsson (nom i cognoms)

amb el títol de: It's Not Whether you Win or Lose: Investigating the Use of Serious Games and L2 Reading Development

certifico que he llegit el treball i l'aprovo perquè pugui ser presentat per a la seva defensa pública.

I perquè consti i tingui els efectes oportuns signo aquest certificat en

Barcelona, a 28 de juliol de 2020

Dr/a. Roger Gilabert Guerrero

UNIVERSITAT DE BARCELONA

Facultat de Filologia i Comunicació
Dept. Llengües i Lit. Modernes i Estudis Anglesos

Gran Via de les Corts Catalanes, 585
08007 Barcelona, SPAIN
Tel. +34 934 035 686
Fax +34 933 171 249

Official MA programme in
Applied Linguistics and Language Acquisition in Multilingual Contexts
(LAALCM)

Universitat de Barcelona

Non-Plagiarism Statement

This form must be completed, dated and signed and must be included at the beginning of every copy of the MA Thesis you submit for assessment.

Name and surnames:	David S. Israelsson
MA Thesis title:	It's Not Whether You Win or Lose: Investigating the Use of Serious Games and L2 Reading Development
Supervisor:	Dr. Roger Gilabert Guerrero

I HEREBY DECLARE THAT:

- This MA Thesis that I am submitting for assessment is entirely my own work and I have written it completely by myself.
- I have not previously submitted this work or any version of it for assessment in any other programme or institution.
- I have not used any other sources or resources than the ones mentioned.
- I have identified and included the source of all facts, ideas, opinions and viewpoints of others through in-text referencing and the relevant sources are all included in the list of references at the end of my work. Direct quotations from books, journal articles, internet sources or any other source whatsoever are acknowledged and the sources cited are identified in the list of references.

I understand that plagiarism and copying are serious offences. In case of proof that this MA Thesis fails to comply with this declaration, either as negligence or as a deliberate act, I understand that the examiner has the right to exclude me from the assessment act and consequently all research activities conducted for this course will be declared null and the MA Thesis will not be presented for public defense, thus obtaining the lowest qualification.

Date: 30 Sept. 2020

Signature: David Israelsson

Acknowledgements

The production of this study has been one of the most challenging and rewarding experiences of my life. Its size and scope greatly eclipse any of my other academic undertakings. Insufficient data forced us to broaden our pool of participants significantly, and a lack of findings led us to narrow our focus. This offered a valuable experience and highlighted that in the real world, plans must be adapted to unforeseen challenges and seldom conform to initial designs. It further imparted on me a great degree of respect and admiration for those who choose to pursue research as a vocation. Attending the University of Barcelona's LAALCM program, researching, conducting and writing this thesis: these are things I will carry with me always.

This study would not have been possible without the immeasurable contribution of Dr. Roger Gilabert Guerrero, whose guidance and advice led to its design and undertaking. It was at his suggestion that we did not accept a null effects outcome and narrowed our focus. Further acknowledgements are owed to Matthew P. Pattemore and Judit Serra whose data collection, knowledge of Tableau, and expertise greatly streamlined and aided in this study. I also want to acknowledge Dr. Elsa Tragant, whose classes clarified expectations regarding sound thesis design and organization. Further thanks are owed to the University of Barcelona LAALCM program, and the committee members overseeing the defense of this thesis. Also, I must express my gratitude to the schools and participants involved in this study, who allowed us to collect data and monitor their game use, as well as those who developed and instituted the iRead program. Lastly, this study would not have been possible without the love, support, and technical know-how of Sarah E. Schefers. I am grateful to you all. This study is dedicated to Venerable Luang Po Doem.

It's Not Whether You Win or Lose:

Investigating the Use of Serious Games and L2 Reading Development

By D. S. Israelsson Supervisor: Dr. Roger Gilabert Guerrero

Abstract

With the reliance on technology becoming increasingly prevalent in the classroom, educational administrators and educators have begun to consider how to effectively incorporate serious games into their curriculums and lessons. Published research supports that early mathematics and reading development are areas in which serious games have proven to be beneficial in both the short and long-term. This longitudinal study investigates the effectiveness and mechanisms involved in the development of L2 reading skills through the use of the novel serious game iRead. Hoover and Gough's (1990) simple view of reading demonstrates that the dual components of word reading accuracy and fluency form a measure of word recognition, which, along with listening comprehension is responsible for explaining differences in reading comprehension. iRead is a EU-funded project that seeks to create adaptive technologies that will contribute to improvements in reading skills. By collecting measures of word reading accuracy and fluency from 72 ESL learners prior to and following four months of iRead use, this study sought to tie overall accuracy and fluency gains to game use and performance. We also consider differences in game use and performance based on initial proficiency measures in listening, reading, and vocabulary. We link gains in word reading accuracy to the use and performance of a specific iRead feature and also show evidence of iRead's adaptivity, drawing conclusions about its role in overall gains and gameplay. Finally, we tie gains in fluency to number of books read and number of tricky words saved during iRead use. This study contributes to the body of existing research investigating the effectiveness of personalized and adaptive serous games and provides evidence for their efficacy when used in conjunction with traditional methods of L2 reading development.

Keywords: Personalized and adaptive games, serious games, L2 reading, the simple view, word reading accuracy, fluency

Table of Contents

1. Introduction

Electronic gaming has evolved from a fringe hobby in the 1970s to a worldwide multi-billion dollar industry today. In 2019, the Entertainment Software Association estimated that three in four Americans live with someone who regularly plays videogames, and that 164 million adults played regularly. This popularity has led to videogame designs evolving from origins in entertainment to becoming platforms for training and education. Serious games, or games whose primary purpose is not entertainment (Susi, Johanneson & Backlund, 2007), continue to grow in complexity and popularity. Game designs have advanced from simple rote-learning tools to highly complex and adaptive systems. Language learning applications such as Duolingo have become popular low-cost educational alternatives to language classes among private consumers. Curriculum designers and institutions, on the other hand, have shown reluctance to the adoption of serious games due to a paucity of research investigating and supporting their effectiveness, and a lack of clarity as to their function and potential as part of the concept of personalization. Thus, the role of serious games as an alternative or complement to regular face-to-face lessons has yet to gain general acceptance.

With the development of adaptive algorithms, there is a greater potential for serious games to address more complex areas of learning, such as the development of reading skills. Within the EU-funded project iRead, Navigo is one such serious game which combined with the Amigo e-reader has been designed to help with and measure reading development.[1] Using an algorithm capable of adapting to learners' performance on games and features, it relies on a built-in linguistic infrastructure containing a domain model that specifies language features, and a large morphologically annotated dictionary on which games draw to extract words. iRead's ability to respond to the needs of individual learners based on input has the potential to support the development of L2 reading. Drawing upon numerous proficiency measures and a pre-test

[1] iRead (infrastructure and integrated tools for personalized learning of reading skills) is a 4-year EU-funded Horizon 2020 'innovation action' whose overarching aim is to develop a software infrastructure of personalized, adaptive technologies for supporting learning and teaching of reading skills. The project has developed three tools which feed on one another and are used in coordination: reading games (Navigo), an e-reader with interactive texts (Amigo) and a Teacher Tool (for teacher supervision). 6 European countries (United Kingdom, Spain, Germany, Romania, Greece and Sweden) participate include both 7-year-old novice readers in English, Greek, german, and Spanish as well as 10-12-year-olds learning to read in English as a foreign language. The projects include over 5000 participants. For more information, please visit https://iread-project.eu/.

post-test design, this exploratory longitudinal study investigates whether and how the technological tools included within the iRead project impact the development of reading accuracy and fluency in the L2.

In order to accomplish this, we will first examine the development of reading skills and the constituent components of word recognition and listening comprehension as outlined in Hoover and Gough's (1990) simple view of reading. We will define the concepts of accuracy and fluency, and justify their inclusion as indicators of reading development. We will discuss Cain's (2015) addition of vocabulary knowledge to the simple view. Next, we will consider the challenges posed by the addition of a second language in explaining the mechanisms of reading development. Also as part of the literature review, we will examine reading development in the context of technology. We will define personalization and adaptively in games and consider how their inclusion may improve the effectiveness of a reading development system and differ from traditional classroom methods. Finally, we will examine the iRead system as it relates to the development of reading skills by presenting our analysis of iRead system use as well as any changes in reading development. The conclusions drawn by this study will contribute to the understanding of the effectiveness of personalized and adaptive games as they relate to L2 reading development by linking gameplay to gains in word reading accuracy and fluency.

2. Definition and Development of Reading Skills

The development of reading skills is an important aspect of second language acquisition (SLA) within the classroom context. L1 reading comprehension involves the interplay of multiple cognitive processes (Alderson, Haapakangas, Huhta, Nieminen, & Ullakonoja, R., 2014; Cain, Oakhill & Bryant, 2004; Hoover & Gough, 1990; Nassaji, 2011). Hoover & Gough (1990) presented a theoretical framework outlining the domains involved in the development of reading comprehension skills in their simple view of reading. The simple view as initially explained by Hoover & Gough (1990) holds that the skills involved fall into two independent and distinct components: that of word-recognition, (referred to by Hoover and Gough as decoding,) and listening comprehension (referred to also as linguistic comprehension.) Under this framework, one's reading ability is the result of one's capacity to read and identify words combined with their ability to comprehend aurally presented text (Cain, 2015). These two domains are demonstrated to build upon and influence one another, and the level of reliance

upon one domain or the other by early readers changes with proficiency level and age (Hoover & Gough, 1990). The conclusions drawn by Hoover and Gough in the simple view have been supported and built upon by further research. Cain (2015) for example, demonstrated that the simple view model explained around 90% of the variance in L1 reading comprehension among 1st through 3rd grade participants (pp. 22). In addition to these two independent domains, Cain (2015) further submits that vocabulary knowledge plays a third distinct and integral role affecting both domains and includes it in a 'new' simple view model. This inclusion plays a particularly important role in the study of second language reading development, where a lack of vocabulary knowledge may negatively affect word recognition and listening comprehension skills (Nassaji, 2011). Despite the importance of vocabulary knowledge to L1 and L2 readers, the path to literacy for beginner-level readers begins with the development of word recognition.

The ability to associate written text on a page with semantic meaning is an important milestone which must be met in order for a young reader's proficiency level to increase (Cain, 2015). Nassaji (2011), notes that the ability to visually obtain information is unique to reading and does not exist in other forms of language such as listening or speaking, though sign language is not considered. Decoding, hereafter referred to as word recognition is defined by Hoover & Gough (1990, p. 130) as "the ability to rapidly derive a representation from printed input that allows access to the appropriate entry in the mental lexicon." Additional definitions of word recognition, such as Gough and Tunmer's (1986), include the adjective 'silently' as an integral component. Cain (2015, p. 7) cautions, however, that including 'silently' in any definition of word recognition makes it "difficult to operationalize or empirically measure." She notes that silently read words cannot be measured for accuracy, and that due to the influence of listening comprehension within the structure of the simple view, comprehension tests are not strict measures of word recognition. Hoover & Gough (1990) demonstrate that beginner-level readers rely heavily on word recognition as they are tasked with decoding words that they may have never seen in written form. Word recognition skill can be predicted by measures of reading accuracy and fluency (Cain, 2015; Hoover & Gough, 1990).

2.1. Word Reading Accuracy

As a component of word recognition, word reading accuracy is the ability of a reader to correctly generate a phonological representation of an encountered lexical item (Cain, 2015).

Beginner-level readers often embrace the time consuming and effortful strategy of sounding words out, especially in languages where orthographies match their phonological representations (Bernhardt & Kamil, 2006). This strategy limits fluency and comprehension due to its lack of prosody and cognitively taxing nature (Perfetti, 2007). As accuracy improves, reading proficiency increases whereby readers begin to instantly recognize words through sight-reading. (Nation & Snowling, 1998; Stanovich & West, 1979) Conversely, sight reading may be less frequent in L2 learners who are exposed to literature as a means of introducing new vocabulary and grammar concepts. Bernhardt and Kamil (2006) note that it will, however, be embraced for frequently encountered lexical items.

Accuracy has been operationalized in numerous ways, including word and non-word decoding tasks. The ability to accurately pronounce printed high and low frequency words within a period of time is also used (Cain, 2015), and is operationalized in this way within this study. As a component of word recognition, measures of accuracy are often used within the simple view of reading framework as reflections of early reading development before fluency can be accurately measured. As readers begin to more effortlessly decode the lexical items they encounter, the speed at which they read increases. Cain (2015) found that 1st graders rely on accuracy far more than 3rd graders by using letter-by-letter sub-lexical strategies which develop progressively into lexical, whole-word strategies that increase reading speed. This finding illustrates that reliance on word reading accuracy diminishes with greater reading proficiency and increasing reading fluency.

2.2. Reading Fluency

Whereas a definition of word reading accuracy has been generally agreed upon by researchers, the same cannot be said for reading fluency which is-a critical aspect in diverse fields including linguistics, SLA, literacy development and even public policy. Cain (2015) notes that it is a complex construct involving a bridge between the ability to decode and the ability to derive meaning. Harris & Hodges (1995, p. 85) define fluency as the "freedom from word identification problems that might hinder comprehension." Fuchs, Fuchs, Hosp and Jenkins (2001, p. 239) define it as "…the oral translation of text with speed and accuracy," while Cain (2015, p. 8) adds that "accurate expression or prosody" should also be included. Within the context of this study, Fuchs et al's (2001) definition of fluency will be adhered to.

As with accuracy, fluency has been operationalized in a number of ways. Because of its nebulous definition, several direct and indirect measures have been devised. Johnston and Kirby (2006), for example, operationalize it indirectly as the time taken to name an array of drawn objects while others have used letters. More direct fluency measures us phonemic decoding tasks which can be comprised of word or non-word lists. These phonemic decoding tasks are used to generate a fluency index by measuring the speed at which one reads by measuring words per minute or seconds per passage (Cain, 2015). As will later be described, this study operationalizes fluency with such a timed reading passage. Cain also notes that some measures of fluency take accuracy into account while others do not, further illustrating the complex relationship between the two.

Measures of fluency and accuracy both serve as indicators of word recognition skill, a critical domain within the simple view of reading. A consensus has not yet been reached as to the relationship between fluency and accuracy. This is due to relatively few studies investigating this relationship as well as contradictory findings (Cain, 2015). More specifically, attempts to isolate fluency from accuracy have yielded results that show that fluency and accuracy are either separate (Cain, 2015; Protopapas, Simos, Sideridis, & Mouzaki, 2012) or interconnected constructs (Adlof, Catts & Little, 2006), offering an important avenue for future study. Cain further finds that word recognition of less proficient readers is better measured by word reading accuracy while fluency is a better measure for more developed readers. Within the SLA context, the use of accuracy and fluency as predictors of reading comprehension may be further complicated by factors such as the orthography of the language being read, the level of transfer between L1 and L2, grammar, and affective measures (Bernhardt & Kamil, 2006; Nassaji, 2011). As will be described in instruments, changes in accuracy and fluency are used as indicators of changes in overall word recognition between pre-test and post-test measures within this study.

2.3. Reading Development: The Simple View of Reading

Listening comprehension is defined as the ability to take aurally transmitted information and derive sentence-and discourse level interpretations. (Gough & Tunmer, 1986; Cain, 2015) Listening comprehension is generally assessed by measuring the ability to answer questions about the content of a spoken or recorded narrative. It may also include the retelling of such narratives. In the assessment of reading comprehension, one's ability to not only recognize

words but also to interpret and reflect upon them is measured illustrating the equal importance of the simple view's two domains. As part of this study, a measure of initial L2 listening proficiency is taken using the Cambridge Preliminary English Test.

When published, Hoover and Gough's (1990) *The Simple View of Reading* made considerable contributions to contemporary notions of literacy, reading disability and reading instruction. This study and accompanying theoretical model demonstrated that "a combination of [word recognition] and listening comprehension made substantial contributions toward explaining variation in reading comprehension" among English-Spanish bilingual children (pp.127). Furthermore, Hoover and Gough's simple view made two influential predictions regarding reading development as well as reading difficulties (Cain, 2015). It predicted that as the developing reader's word recognition skills increase, the relative weightings of word recognition and listening comprehension would change (Gough, Hoover & Peterson, 1996). The simple view also predicted that difficulty in reading comprehension may be the result of either poor word recognition, poor listening comprehension, or a combination of the two (Gough & Tunmer, 1986). Cain (2015) notes that there is broad support for the simple view's predictions, and that cross-sectional studies show that as reading proficiency increases, the influence that listening comprehension has on reading comprehension increases. This is a result of novice readers moving from reading letter by letter to reading whole words with increasing accuracy and fluency, lending lesser weight to the importance of word recognition. The simple view's predictions have been replicated over time. For example, Cain's (2015) study found that approximately 90% of the variance in reading comprehension was explained by these two domains among first, second, and third grade U.S. students. Absent from the simple view of reading model, however, is the level of influence vocabulary knowledge may have on word recognition and listening comprehension.

Despite being published 30 years ago, the simple view of reading has stood up remarkably well to attempts to amend or restructure it. One area of current research is the extent to which individual differences in vocabulary knowledge affect word recognition and listening comprehension. In her 2015 review of the simple view of reading, Cain found that vocabulary indirectly predicted reading comprehension by influencing both listening comprehension and word recognition independently. Other studies have partially supported this, finding that

individual differences in vocabulary can predict listening comprehension skills (Nation & Snowling, 2004; Ouellette & Beers, 2010) or word reading ability individually (Mitchell & Brady, 2013; Nation & Snowling, 2004; Ouellette, 2006). Cain submits an amendment to the simple view's model in which vocabulary independently influences both of the domains comprising reading comprehension. This is of particular interest in the study of L2 reading skills because limitations in L2 vocabulary knowledge may force learners to rely more heavily on simpler strategies that had been abandoned in L1 reading.

2.4 Reading in a Second Language

As reading proficiency increases, studies show that readers' reliance on specific domains change over time. (Cain, 2015; Hoover & Gough, 1990.) Novice readers move from a sub-lexical strategy (of reading letter by letter) to a lexical one where they begin to consider whole words. Lexical strategies have a clearly positive impact on reading fluency and accuracy. As for reading in the L2, students who may already be using lexical strategies in their L1 may revert to sub-lexical strategies for their L2. Furthermore, their lexicons in the L2 may be so limited that they do not facilitate the use of lexical strategies whatsoever.

Many models that seek to explain L2 literacy are grounded in and extensions of L1 research. (Nassaji, 2011). Reading is often a necessary and integral skill in the acquisition of a second language, however the addition of an L2 complicates the theoretical framework of the processes at play due to numerous factors (Barnhardt & Kamil, 2006). One reason is that L2 reading involves an interaction between 2 languages whereas L1 reading involves only one (Koda, 2007). Another is that reading comprehension in the L2 may be negatively impacted by limitations in vocabulary that do not exist in the L1. Many L2 learners begin reading in the L2 without the extensive vocabulary required to extract sufficient meaning from the text. For these learners, reading is a tool for developing basic language skills and not for extracting deeper meaning. (Nassaji, 2011) Another reason stems from the level of transfer that may exist between the L1 and the L2 (Bernnhardt & Kamil, 2006). Due to orthographic, phonological, syntactic, morphological and semantic differences, reading processes differ depending on language (Grabe, 2009). Grabe (2009) further notes that linguistic differences impact early L2 readers at a greater level than more proficient readers. Finally, the level of reading proficiency in the L1 may

positively or negatively affect the proficiency in the L2 in line with Cummin's *Linguistic Threshold Hypothesis* (Bernnhard & Kamil, 2006; Brevik, Olsen & Hellekjär, 2016).

3. Reading and Technology

The introduction of personalized and adaptive learning technologies has facilitated the transformation of traditional 'cookie cutter' teaching methodologies into lessons that are able to respond to the individual differences of learners (Gomez, Zervas, Sampson & Fabregat, 2014). Wu, Chang, Chang, Liu & Heh (2008, p. 96) define personalized and adaptive educational tools together as, "the process of enabling the system to fit its behavior and functionalities to the educational needs (such as learning goals and interests), the personal characteristics (such as learning styles and different prior knowledge), and the particular circumstances (such as the current location and movements in the environment) of the individual learner or group of interconnected learners." Within this relatively broad definition, a technology can be personalized, or adaptive, or both.

The definition of what it means to be a 'personalized technology' remains a topic of debate depending on context (Holmes, Anastopoulou, Schaumburg & Mavrikis, 2018). As with reading fluency, personalized technologies are not a field-specific construct and thus what may constitute an adequate definition within one field may not suffice for another. In the SLA context, personalized technologies are ones that are capable of personalized learning. Holmes et al (2018) view personalized learning as an extension of personalized teaching strategies embraced by teachers who may focus on one topic or another depending on the needs of the learner. They note that a common definition of personalized learning may be identified by its features. It seeks "to improve student engagement and achievement, focus on meeting individual learning needs, shift to changing needs as well as recognize that individuals progress at different rates" (p. 17).

Adaptivity is the ability of a system to take into account specific learner characteristics in order to design an appropriate learning experience (Martin & Carro, 2009). It can be viewed as an instantiation of personalization. This emergent technology requires a system with the ability not only to gather and save information from its user in real time, but also an algorithm that is able to continually adjust to new input and to amend its output according to contextual

information (Gomez, Zervas, Sampson & Fabregat, 2014.) Within the context of literacy development, adaptivity is useful because it provides individual feedback and support to learners in the form of scaffolding. It allows learners to notice and learn at an individual pace, introducing them to new material while forcing them to revisit concepts that are not demonstrably mastered. This upends traditional teaching methods in which groups of students are made to keep up with the pace of the curriculum. As personalized and adaptive serious games have grown in popularity and entered the classroom, their effectiveness as learning tools has become a topic of closer study. In the present study, the iRead system uses an adaptive algorithm in order to respond to learners' individual paces and abilities and will be described at length in the instruments section.

While traditional reading development has been tied to the amount of reading exposure via text, there are lingering questions about whether this applies equally to digital environments. Deligiannis, Panagiotopoulos, Patsilinakos, Raftopoulou, and Symvonis, (2019) illustrate the potential strengths of personalized and adaptive serious games by comparing them to learning activity books. They note that activity books are of limited size, static, and cannot be revised. Parents and teachers must anticipate the proficiency of the learner and ensure that the activity book is of a commensurate skill level. Activity books are impersonal, and their content does not change from viewing to viewing, limiting their utility. Alternatively, the adaptive nature of serious games allows them to continuously adjust to the individual skill level of the learner. They are personalized and adaptive in that the material presented takes into consideration whether previous material was 'learned' or not by a specific learner. Serious games are also dynamic, with their interactive nature allowing unlearned concepts to be revisited at a later time in a different format or activity.

Serious games have become increasingly popular educational and training tools as the technology enabling their use has developed (Michael & Chen, 2006). Questions about their suitability to facilitating literacy and mathematics development have led to a small but growing corpus of research in recent years. Connolly, Boyle, MacArthur, Hainey, and Boyle, (2012), note that digital (serious) games are especially suited to reading and mathematics development because they are "experiential, situated, problem-based and provide immediate feedback" (p. 661). However, Vanbecelaere, Van den Berghe, Cornillie, Sasanguie, Reynvoet, and Depaepe

(2020) remark that empirical evidence is mixed regarding the effectiveness of (serious) games and reading development, especially in the long term. For example, they found that among 336 primary school children in Flanders, participants who played a reading game improved significantly more in simple word reading fluency and text reading fluency than those who did not immediately following intervention. However, they found no such difference for complex word reading fluency. In a delayed post-test, they found that students who used the reading game outperformed students who did not in simple word reading fluency, but not in text reading fluency. Furthermore, they found no difference in calculation fluency among participants who used the mathematics-based Number Sense game and participants who did not immediately following intervention. This study is valuable in that it shows that serious games can be beneficial to certain aspects of reading development in both the short and long-term.

Other studies echo the short and long-term benefits to reading development from serious games. For example, Kartal and Terziyan (2016) found that among 20 low-SES participants, those exposed to a serious game focusing on phonological awareness were able to outperform those that were not in letter-naming and phoneme segmentation, but not on rhyming or syllable blending. van de Ven, de Leeuw, van Weerdenburg, and Steenbeek-Planting (2017) examined early literacy skills among 60 children with special needs through the use of the serious game Letter Prince. They found immediate and delayed effects on text reading fluency in favor of those in the serious game condition, and immediate enhanced effects in pseudo-word reading fluency, but no effect on decoding. Among a control group and groups of students playing serious games focusing either on rhyming skills or letter-sound correspondences, Kyle, Kujala, Richardson, Lyytinen and Goswami (2013) found that participants who played serious games outperformed those who did not in immediate and delayed reading, spelling and phonological skills posttests. Generally, these studies examine participants who are young and learning to read in their respective L1. Older learners, with greater variance in proficiency level, may experience greater effects from a games' adaptivity. Also, few studies have investigated the effects of serious games on L2 reading development, illustrating a gap in the literature.

In sum, from the extensive literature on reading and the still scarce literature on reading through technology we have distilled a number of gaps that need to be addressed. Firstly, reading accuracy and fluency seem to improve with reading practice, but it has yet to be shown whether

this applies to reading in the context of serious games. Secondly, an issue that has only been minimally addressed is whether work on specific linguistic features that are involved in reading (e.g. grapheme-phoneme correspondence) may actually contribute to overall improvements in accuracy and fluency in the context of serious games. Thirdly, more evidence is needed about the role of decoding, listening, and vocabulary and their contribution to L2 reading development since to date no longitudinal study has incorporated the simple view of reading in relation to L2 accuracy and fluency gains through personalized and adaptive serious games. Because of the exploratory nature, no directional hypothesis will be proposed. Our focus is on measuring the impact of using iRead's adaptive system on the development of reading skills as measured by word reading accuracy and fluency gains.

RQ1: Does playing a larger number of games impact overall gains in reading accuracy and fluency?

RQ2: Does a better performance on the games explain gains in the L2 reading accuracy and fluency?

Research question 1 is distilled from the literature on reading where a central idea is that reading skills increase with reading practice, but does engagement with a larger quantity of digital games that focus on common reading issues improve overall reading accuracy and fluency? As for research question 2, in this study we wonder whether getting more correct answers in the games leads to improved reading skills since from the literature it is also known that improving performance on specific linguistic features may lead to overall reading gains.

4. Methodology

In order to answer the research questions above, this exploratory study used a quasi-experimental pretest/posttest design in which learners' reading skills were tested before the use of iRead and after 3 to 4 months of use of the system on a weekly basis. System use occurred in class individually, but under teacher supervision. Below we first describe the participants. We go on to give a detailed explanation of the instruments and justify their use.

4.1 Participants

The participants for the present study were part of a larger longitudinal investigation into the effectiveness of iRead as a tool for first and second language reading development involving 6 E.U. member states and approximately 6200 learners ranging between 5 and 12 years of age. Participants in this study were enrolled at five of the seven participating primary schools in Catalonia; four large primary schools (1 free school in a large city, 3 free schools in medium-size cities) and the fifth a smaller and more rural school. They were all learning English as a foreign language in regular classes and, as per agreement with schools, used the iRead system (games and reader described below) for one hour every week during a period that ranged from 3 to 4 months. Due to the 2020 mandatory school closures, there was considerable attrition among our participants, particularly with respect to obtaining post-test data.

Table 1: Participants and exclusions by school and gender.

	Male Participants		*Female Participants*		*Total Participants*	
School	Males (Excluded)	Remaining	Females (Excluded)	Remaining	Total (Excluded)	Remaining
1	11 (6)	5	5 (3)	2	16 (9)	7
2	27 (24)	3	27 (19)	8	54 (43)	11
3	24 (18)	6	23 (19)	4	47 (37)	10
4	33 (23)	10	19 (11)	8	52(34)	18
5	31 (14)	17	24 (15)	9	55(29)	26
Total	126 (85)	41	98 (67)	31	224 (152)	N = 72

Of the original participants (N = 224), 152 were excluded due to incomplete data. Many of these exclusions were the result of voluntary post-test submission. This led to numerous cases of posttests being mislabeled, improperly recorded or simply not submitted. Other exclusions resulted from technical difficulties, absences during pre-testing, or non-use of iRead's adaptivity feature. The remaining participants, (N = 72, 41 males and 31 females) completed all tests and were 6th year primary school students who as per the Spanish curriculum had been studying English for 5 years. No information was obtained about their familiarity with tablets or games.

4.2 The iRead games[2]

[2] For a complete description of the iRead project, please visit https://iread-project.eu/

During weekly iRead class, learners were encouraged to play both the games and use the e-reader[3] in the tablets supplied by the iRead project. Games were included in the Navigo app (developed within the iRead consortium by professional game developers *Fish in a bottle*) installed on Android tablets. Each student had an account with a username and password, and was encouraged to create his or her own avatar, a crucial element in gameplay to generate and maintain engagement. The games were contextualized in an Egyptian setting where the simple narrative instructed them to get into the 'Pyramid of Lost Words' in order to save the inhabitants of an oasis who had hidden in the pyramid during a storm. All games were designed to consider language choices in order to solve mini language puzzles. The games included 13 different mechanics[4] (e.g. crossing a bridge by choosing linguistic options in *Bridgytian*, slicing the syllables of words in *Slycecophagus*, or working out morphological problems in *Crocotiles)* which were associated with the different features of the English language, generating thousands of combinations. The use of the games was overwhelmingly greater than the reading of texts, since the e-reader app (Amigo) displayed some problems. For reasons of space, the Amigo app is not described in any detail here. It suffices to say that it includes texts and a voice system that reads each text so that learners can connect word forms to their phonological form. Learners in the five schools played a total of 24,327 games during the 3-4 month period, which covered a total of 1,018 features. Participants also read a total 4041 stories and recorded 923 tricky words.

From a linguistic point of view the games covered features from the English language which are relevant to the development of reading. 279 features from the phonological (e.g. grapheme-phoneme correspondence or syllabification among others), orthographical (e.g. confusing letters), word recognition (e.g. sight words), morphological (e.g. prefixes and suffixes) morphosyntactic (e.g. verb endings) and syntactic levels (e.g. relative clauses) which were

[3] The e-reader (Amigo) included hundreds of texts in English selected according to the different features they contained and were classified into different levels by reading experts at the University College London (UCL). Texts that contained the features the readers were playing on the Navigo app were selected for reading. Each text included a pre-reading explanation about a feature (e.g. the fact that in English the sound /s/ can correspond to more than one letter – e.g. '-s', '-ss', '-ce', …). Then readers could read the text silently as they listened to an automatized voice that read the text for them (they used headphones for both the games and the e-reader app). The feature included in the explanation was also highlighted in the text. Among other features, readers could click on a word and get its meaning and pronunciation, and they could create their own personalized list of 'tricky words'. In typical iRead lessons, learners would play 5 to 15 games and read one or two short texts.

[4] 1.Cleomatchra (Pairs), 2.Bridgyptian, 3.Hearoglyphs, 4.Raft Rapid Fire, 5.Pillar Pusher, 6.Walk Like an Egyptian, 7.Remove the Runes, 8.Slicecophagus, 9.Cogelisk, 10.Perilous Paths, 11.Croco-tiles, 12.Anubrick, 13.Cartastrophe

included in the *domain model*, a system-internal tool that specified the characteristics of each feature, and predicted level of difficulty for each of the features, among other technical information. Most importantly, the domain model specified the pre-requisites for each feature (e.g. diphthong [au] was played once individual vowel sounds [a] and [u] were mastered). Another important components of the iRead infrastructure included a 17 thousand word *dictionary* that specified the orthographical, phonological, morphological and syntactic information of each word among others. Both tools were devised by the University College London teams in the iRead team. The tools were used to provide content for the games.

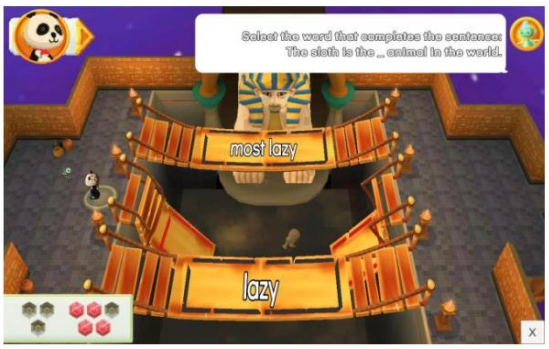

Figure 1. Example of *Bridegyptian* game with three morphosyntactic options

From a technical point of view, the games followed an adaptive logic. Learners started at a given point, determined by their year at school, with a limited number of opened features, and they were offered the same or different features on the basis of their performance. An adaptivity algorithm computed the leaner's performance on each mini-game (number of successes and failures) and adapted to the pace and performance of each learner. When learners successfully completed the same feature three times, they would advance to another more difficult feature. If they failed, students were offered more and different mini-games on the same features until it was mastered. Mastery of any one feature was set at 70%. The long experience of the game developer in the iRead team guaranteed high levels of interest and engagement through action, variety, emotion, music, colors, rewards, and all kinds of game elements that kept learners focused on solving the mini-games. In doing so, they had to necessarily go through the consideration, comparison and contrast of the features at play and this was meant to raise their linguistic awareness of the features. For a high percentage of the features some kind of pre-

recorded feedback was provided, either in the format of outcome feedback (e.g. green highlight when right and red when wrong), and sometimes some pre-recording emotional (e.g. 'well done!' if they got it correct or 'keep trying' if they failed), and/or elaborative feedback (e.g. double "ss" is pronounced [s] as in 'glass') (Pattemore, Gilabert & Serra, 2019). Because of the adaptivity algorithm and the pre-requisites specified in the domain model, during the 3-4 month period EFL learners primarily worked at the phonological level, although learners also played with a few morphological and syntactic features.

4.3. Instruments

Both system-external and system-internal instruments were used in order to answer the research questions for this study. System-external instruments tapped on leaner's individual differences in reading skills were used at pre-test and posttest. The test included a list of L2 words for measuring reading accuracy and a list of L2 non-words for the measurement of reading accuracy both developed by UCL, and a reading fluency test taken from the FAIR-FS test (Foorman, Petscher, & Schatschneider, 2015). System-internal data including the number of games and number of features per game among others were collected by means of the learning analytics embedded in the iRead infrastructure and were extracted via Tableau[5].

4.3.1. L2 Word Reading Accuracy

An L2 word-reading accuracy measure (WA) was taken from participants as part of the pre-test (T1) and post-test (T2). This task was part of a longer battery of tests which also included measures of L2 non-word reading accuracy, L2 fluency, L1 fluency, and working memory. It was comprised of a list of 90 words (See Appendix A), beginning with high frequency monosyllabic words and progressing to lower frequency and phonologically complex words. Participants were instructed to read as many of the words as possible in one minute as clearly and accurately as possible, and recorded. These recordings were analyzed by a native English speaker noting each correct and incorrect response. An accuracy score was calculated by subtracting the total number of inaccurately read words from the total number of words read in one minute (maximum 90.) Pre-tests were conducted at the participants' schools in September and October of 2019. Post-tests were voluntarily recorded and submitted by participants using

[5] Tableau is an interactive data visualization software. For more information please visit https://www.tableau.com

their own recording media and a list of the words provided by the researchers in April and May of 2020. An index of gains was calculated by subtracting individual pre-test scores from post-test scores, generating 3 measures in total: WA-T1, WA-T2, and WA gains.

4.3.2. L2 Non-word Reading Accuracy

A Non-word reading accuracy measure (NWA) was conducted immediately following the WA measure on the same recording. It was comprised of a list of 66 non-words beginning with monosyllabic two letter words increasing in complexity and length (See Appendix B). Participants were told that they would have one minute to read as many of the words as they could as clearly and quickly as possible. They were instructed to pronounce the words as they thought they should be pronounced if they were English words. A native English speaking researcher listened to the recordings and made note of the items as correctly or incorrectly pronounced according to the conventions of Standard American and British English pronunciation, and following an outline of specifically targeted features. Again, post-tests were voluntarily submitted through a Google survey in the same way as the WA post-tests. Participants were instructed not to prepare for the post-test in any way, but simply to look at the words and read them as they had done on the pre-tests prior to the introduction of iRead. A NWA score was calculated by subtracting the number of incorrectly pronounced items from the total number of items read in one minute (maximum 66.) Gains were again calculated by subtracting pre-test scores from post-test scores. In sum, 3 measures were generated: NWA-T1, NWA-T2, and NWA gains.

4.3.3. L2 Reading Fluency

A reading fluency measure was collected after the non-word reading accuracy measure on the same recording. Participants were informed in English by script that they would be reading from a brief passage entitled “How to make Play Dough” (See Appendix C). The entire passage contained 287 words of various frequency levels. Researchers instructed the participants in English to read as clearly and quickly as possible and that they would have one minute as timed by stopwatch. A native English speaker analyzed the recordings and made note of inaccurately pronounced and missed words. The total number of inaccurately pronounced or missed words were subtracted from the total number of words read in one minute in order to calculate a reading fluency score in words per minute (WPM.) As with the other tests, pre-tests

were conducted in September and October of 2019 on site. They were recorded directly following the NWA pre-tests. Post-tests were recorded independently and submitted voluntarily. During the post-test, participants were given access to a shortened version of the same passage which included only 109 words. In cases where participants read the passage in under one minute, a WPM score was calculated by dividing the number of correctly read words by the number of seconds it took to read the passage and multiplying the quotient by 60. Again, gains were calculated by subtracting T1 scores from T2 scores, generating 3 measures in total: WPM-T1, WPM-T2, and WPM gains.

4.3.4. L2 Proficiency Measures

Measures of L2 proficiency were also conducted with participants in September and October, prior to their introduction to iRead. These were administered on-site at participants' schools by the same researchers administering the pre-tests. Scores obtained from these tests were used to answer questions pertaining to overall L2 proficiency. L2 Proficiency measures included the listening and reading section of the Preliminary English Test (PET) designed by Cambridge University. The PET's listening and reading sections were administered to students on two different days. Overall performance scores and percentages of scores were obtained for listening comprehension and reading comprehension. Based on individual participant scores, a Common European Framework of Reference for Languages (CEFR) language rating was assigned to each participant. All participants received a categorical rating between A1 and B1.

The Picture Vocabulary Size Test (PVST) was used as a measure of receptive L2 vocabulary size. This test was developed by Nation and was initially used to measure the receptive vocabulary size of pre-literate L1 speakers (Nation & Anthony, 2016)[6]. The test has also been shown to be applicable in the L2 context, where participants' reading abilities may be highly varied or non-existent. The PVST is comprised of a series of panels of four pictures whereby participants are asked to identify one of the pictures as described by a word. Vocabulary scores were calculated for individual participants based on the number of correct responses.

4.3.5. iRead Application Measures

[6] Adaptation provided by Eva Puimège from the Catholic University at Leuven to which we are enormously grateful.

As mentioned above, participants were supplied with individual tablets on which they used the iRead application for approximately one hour per week under teacher supervision. All actions of the participants' gameplay were recorded by the iRead infrastructure. Data collected by the game that is pertinent to this study included total number of games played, total number of features played, and outcome of games played (with an end-state of 'success,' 'failure,' or 'quit). Using these data, we were able to calculate a rate of games per feature, and a success rate. Also used in this study were data pertaining to number of books read, number of tricky words saved, and performance on features containing *kn-*, *st-*, *spl-*, *-o_e*, and *u*, which corresponded to items on the word reading accuracy measure.

4.4. Procedure

Pretest and L2 proficiency measures as well as other measures not described here (working memory test, attention switching test, inhibition test, and L1 fluency) were administered by trained research staff in quiet classrooms within the participants' schools in two pre-test sessions. Researchers individually recorded and timed participants undertaking the word-reading accuracy measure, the non-word accuracy measure, and the reading fluency measure. Examiners used answer booklets to make note of biographical data and take any pertinent notes during the pre-tests. Individual pretests (word recognition, non-word recognition, and fluency) took approximately 7-10 minutes to complete. Recordings were timed and catalogued online for analysis, and examiner answer booklets were archived for later reference if needed.

Due to the closure of all primary schools in Catalonia in March of 2020, the procedure for obtaining post-test data was altered. In late April, schools were provided with the word, non-word and fluency test materials and asked if they wished to continue participating in the study. Schools that agreed emailed individual participants with the materials and instructions on how to complete the post-tests, record them and submit them to the researchers. The WA and NWA lists were identical to the pre-tests. As previously mentioned, the text for the WPM measure was the same, but shortened to 109 words, leading to an increase in participants completing the task in under one minute. This, in turn, led to a higher number of fluency measures being calculated using the equation mentioned in the instruments section. Post-tests for WA, NWA and WPM measures were timed by researchers during analysis instead of at the time of testing. Again, a

stop-watch was used to time one minute for all three measures. Due to some confusion as to how to label the voluntarily submitted post-test recordings, a large number of received post-tests were impossible to accurately identify leading to a regrettably high level of exclusions.

4.5 Statistical Procedures and Preliminary Analyses

4.5.1. Descriptives

Descriptive statistics can be found for all measures in Appendix D. In cases of measures containing outliers, these outliers were winsorized, having their values replaced with the nearest normal value. No measure contained more than 2 outliers. In cases where data was not evenly distributed, non-parametric tests were used. Spearman correlations and were used to answer RQ1 and RQ2 since some of the variables included in the correlations were not evenly distributed. Where Spearman correlations are used, they are designated as 'r_s =', while Pearson correlations are designated as 'r ='. Finally, Wilcoxon signed ranked tests were used to compare means between variables in RQ1.

4.5.2. Preliminary Analyses

Here we summarize a number of preliminary analyses (see Appendix E for a complete report) conducted before the analysis of the research questions was addressed and findings suggest that:

1) Because of the moderate to strong correlations between pre-test and post-test and among themselves, the tests used in this study were effective indicators of individual participants' word and non-word reading accuracy and fluency over time.

2) The WA, NWA and WPM measures were well-designed, evenly weighted and generalizable.

3) Positive correlations among the different proficiency tests speak to the validity of the PET and PVST as well as the CEFR language proficiency rubric as generalizable indicators of overall proficiency.

4) Positive correlations between the T1 and T2 scores for WA, NWA, and WPM and the proficiency measures obtained from the PET and PVST not only speak to the validity of

the individual measures involved, but also support the simple view's theoretical framework, which will be discussed at length in the discussion section.

5. Analysis/Results

5.1. Research Question 1: Correlations between game use and reading gains

In order to explore the relationship between game use and reading gains, data were obtained and calculated from the iRead infrastructure for the former and the pre-test and post-test for the latter. Game use measures included total number of games, total number of features, and number of games per feature.

Table 2: Descriptive statistics of both system-external (accuracy and fluency) and system-internal measures (gameplay).

Measure:	*N*	*Min.*	*Max.*	*M*	*SD*
WA-T1	72	20	69	45.70	10.65
WA-T2	72	25	78	51.00	10.73
WA Gains	72	-13	26	5.89	9.10
NWA-T1	72	26	60	45.42	9.60
NWA-T2	72	30	64	47.06	7.47
NWA Gains	72	-16	18	1.51	7.44
WPM-T1	72	40	144	89.71	21.80
WPM T2	72	57	148	103.97	20.18
WPM Gains	72	-15	42	14.92	13.63
Numb. Games	72	28	227	120.74	44.76
Numb. Features	72	21	72	47.82	11.86
Games/Feature	72	1.21	3.55	2.44	.59

Note: WA-T1 = word accuracy at pre-test; WA-T2 = word accuracy at posttest; WA Gains = gains in word accuracy from pre-test to posttest; NWA-T1 = non-word accuracy at pre-test; NWA-T2 = non-word accuracy at posttest; NWA Gains = gains in non-word accuracy from pre-test to posttest; WPM-T1 = fluency at pre-test; WPM-T2 = fluency at posttest; WPM Gains = gains in fluency from pre-test to posttest; Numb. Games = total number of games played; Numb. Features = total number of features played; Games/Feature = mean number of games played for each feature

The first point to check was whether any gains took place between pre-tests and post-tests. Results of the Wilcoxon Signed Rank Test comparing pre-test and post-test scores suggest that learners improved somewhat uniformly at reading accuracy and fluency, as shown by the

significant difference between T1 and T2 for WA and WPM. NWA also showed a trend (0.06) but was not significant.

Table 3: Results of the Wilcoxon Signed Rank Test comparing pre-test and post-test scores

	N =	*Z* =	*p* =
WA-T1 – WA-T2	72	149.429	< .001
NWA-T1 – NWA-T2	72	156.340	.060
WPM-T1 – WPM-T2	72	174.478	< .001

Note: WA-T1 = word accuracy at pre-test; WA-T2 = word accuracy at posttest; NWA-T1 = non-word accuracy at pre-test; NWA-T2 = non-word accuracy at posttest; WPM-T1 = fluency at pre-test; WPM-T2 = fluency at posttest

The second point to address was whether those gains were attributable to the number and types of games learners played. Spearman correlations were run between total number of games, feature types and ratio of games to features, on one hand, and gains in WA, NWA, and WPM on the other. Moderate positive linear relationships exist between gains in NWA and WPM (r_s = .331, p = .002), WA and WPM (r_s = .392, p < .001) and WA and NWA (r_s = .388, p < .001). This indicates that during the time of study, participants whose reading accuracy improved also experienced a commensurate level of improvement in reading fluency for words and non-words.

The first research question seeks to examine the relationship between these gains and the amount of games played. Can more gains be explained by greater game use? In order to answer the first research question, one-tailed correlations were run between the three measures for gains and the three indicators of game use: total number of games, total number of features, and number of features per game.

Table 4: Correlations (one-tailed) between total number of games, number of features and ratio of mean games per features and gains in word recognition, non-word recognition, and words per minute.

Dependent Variable		*WA Gains*	*NWA Gains*	*WPM Gains*
Number of Games	r_s =	.004	.025	.057
	p =	.485	.418	.318
	N =	72	72	72
Number of Features	r_s =	.001	.055	.022
	p =	.495	.322	.428
	N =	72	72	72
Games/Feature	r_s =	.057	-.012	.112

p =	.319	.460	.175
N =	72	72	72

Note: WA Gains = gains in word accuracy from pre-test to posttest; NWA Gains= gains in non-word accurace from pre-test to posttest; WPM Gains = gains in fluency from pre-test to posttest; Number of Games = total number of games played; Number of Features = total number of features played; Games/Feature = mean number of games per feature

A relationship between number of games played and overall gains would indicate that participants who spent more time using the iRead application experienced greater benefit to their L2 word recognition and reading fluency regardless of number of features. As can be seen in Table 4, no statistically significant relationships were found between overall number of games played and any of the gains measures. This indicates that total number of games played did not contribute to overall gains in WA, NWA and WPM.

A positive relationship between total number of features and any of the gains measures would indicate that participants who encountered a greater number of total features benefitted more than those who encountered fewer overall features. Again, no statistically significant correlations were found between any of the gains measures and number of features. This indicates that overall gains cannot be explained by the overall number of features encountered during game use.

A one-tailed correlation was then run between the accuracy and fluency gains measures and the measure of games per feature. This investigates whether participants that were more successful tended to experience greater fluency and accuracy gains, or vice-versa. Again, no statistically significant relationships were found between any of the gains measures and features per game. Addressing the first research question, these findings show that overall game use cannot explain the level of gains for individual participants, indicating other factors may be responsible for explaining gains in reading accuracy and fluency.

5.2. Research Question 2: Correlations between performance and gains

In order to answer RQ2, participants' game performance was initially addressed using two measures. One measure indicating successful game-performance is the number of games per feature mentioned above. (Typically more successful learners will play the same feature fewer times and move on to new features, while less successful learners will keep playing games on the same feature until it is mastered, hence getting to play a smaller variety of features). As no relationship was found to exist between games per feature and overall gains, this was not useful

for answering the second research question. The success-rate was the other measure for game-performance. The success-rate was calculated by dividing the number of games that ended in 'successes by the square of the number of overall games played. (This method was chosen because of large differences in number of games played.) A positive linear relationship between success-rate and any of the gains measures would show that more successful participants experienced improvements in WA, NWA and/or WPM. Conversely, a negative linear relationship between success-rate and any of the gains measures would indicate that iRead was more beneficial for participants who needed to repeat a feature multiple times until learned. A one-tailed correlation found that no relationship exists between success-rate and any of the overall gains measures. This shows that overall gains in WA, NWA, and WPM cannot be explained by the in-game performance of participants as measured by success rate.

Because both success rate and games per feature did not yield any significant relationships, we then decided to take a closer look at individual features and gains in the corresponding items from the WA measure. As will be discussed further in both the discussion and strengths and limitations sections, this is because participants may not have played many of the features in the games that were tested by the pre- and posttests.

Features containing *kn-*, *u*, *spl-*, *-o_e*, and *st-* were examined because a majority of participants encountered these features and they corresponded to one or multiple items on the WA measure. A moderate positive linear relationship ($r = .353$, $p = .008$) was found to exist between number of times the feature containing *st-* was played and corresponding gains on the WA measure's *st-* items. This indicates that participants who were forced to revisit the feature more often benefitted by experiencing greater WA gains which may point to a positive effect of adaptivity.

5.3. Correlations between vocabulary size, listening and reading ability, game performance and reading fluency gains.

Because of a lack of clear relationships explaining gains through gameplay, we then looked for relationships between overall proficiency measures in L2 listening, L2 reading, L2 vocabulary, and CEFR level, measures of gameplay, and WPM gains. As previously illustrated, moderate positive linear relationships have been established to exist between overall proficiency measures (PET and PSVT) as well as the three measures of gains (gains in WA, NWA, and

WPM). First, one-tailed correlations were conducted between listening, reading, vocabulary and CEFR level and rate of games per feature in order to see if initial proficiency influenced how games were played. A negative linear relationship would indicate that participants with higher overall proficiency scores were able to successfully complete features with fewer attempts. Relationships were indeed negative, but all but one were not statistically significant in strength. A weak negative linear relationship was shown to exist between CEFR level and games per feature (r_s = -.220, p = .032) indicating that participants who achieved a higher level of English proficiency according to the CEFR rubric required fewer attempts to successfully complete individual features, which may be expected.

A one-tailed correlation was then conducted examining the relationship between number of total features and the proficiency measures. A positive linear relationship between total number of features and any of the proficiency measures would indicate that higher proficiency participants were exposed to a higher number of features. This would show that higher proficiency participants were able to complete features more successfully thus advancing further than lower proficiency students. A weak positive linear correlation was found between reading proficiency and number of features (r_s = .291, p < .007). This illustrates that participants who were able to read more proficiently were able to unlock and access a larger number of features.

Next, a one-tailed correlation examining overall proficiency measures and success-rate was performed. A positive linear relationship between any of the proficiency measures and success rate would indicate that higher-proficiency participants were able to successfully complete features more often than lower-proficiency participants. The test showed no statistically significant relationships existing between any proficiency measure and success-rate.

Then, relationships between overall proficiency scores and gains were examined. First, a one-tailed correlation was performed between gains for WA, NWA and WPM and reading proficiency. Next, we looked for relationships between gains and listening proficiency, vocabulary and CEFR level. However, no statistically significant relationships were found to exist between any of the proficiency measures and WA, NWA and WPM gains.

The lack of relationships above led us to take a closer look by running one-tailed correlations between the proficiency measures and gains in five individual features: *kn-*, *u-*, *spl-*, *-o_e*, and *st-*. Again this was the result of the accuracy measure being rather general and not

aligning very well with the features that our participants encountered during gameplay. Of the five individual feature measures, only *st-* produced statistically significant results. *St*-gains showed moderate negative linear relationships with reading proficiency (r_s = -.400, p = .003), and CEFR (r_s = -.369, p = .006). Taken in conjunction with the previously mentioned relationship between number of *st-* games and gains, this indicates that participants with lower reading proficiency and CEFR ratings experienced greater gains on this item due to repeating the same feature more often, another indicator of adaptivity.

Aside from the above-mentioned moderate relationship, weak correlations between CEFR and games per feature, and reading proficiency and number of features, there are few statistically significant relationships to be found between initial proficiency, game use and gains. The lack of strong significant relationships found points to two possible conclusions: either an external, unaccounted-for factor is responsible for gains (such as individual differences, for example), or an internal (within-game) factor that was not included in our measures played a role (such as the role of adaptivity). These possibilities will be discussed at greater length in the discussion section.

5.4. Further Analysis

Because of the exploratory nature of this study, the large amount of data collected and the relative paucity of significant relationships found that could aid in answering our research questions, we decided to conduct further (one-tailed) tests relating to the number of Amigo books read, tricky words saved, and participant gender. Number of books read showed moderate linear relationships with number of tricky words saved (r_s = .360, p = .001), and listening proficiency (r_s = .381, p < .001) as well as weak positive linear relationships with WPM gains (r_s = .207, p = .040) and reading proficiency (r = .200, p = .046). This shows that participants who were inclined to read more were also more likely to make note of unknown words when encountered. It also sheds light on our second research question by illustrating that participants who read more books experienced greater fluency gains. Finally, it supports the simple view by tying listening proficiency to reading.

Number of tricky words saved yielded a number of additional weak relationships, including with gender (r = .205, p = .042), WA gains (r = .277, p = .009), NWA gains (r = .226, p = .028), and WPM gains (r = .240, p = .021). This indicates that participants who were more likely to

highlight unknown words were more likely to experience gains in both accuracy and fluency than participants who were not. Furthermore, weak negative relationships were found to exist between number of tricky words saved and number of games played (r_s = -.227, p = .027), games per feature (r_s = -.223, p = .025), and success rate (r_s = -.210, p = .038) showing that participants who saved more tricky words tended to play fewer overall games and features, and repeat fewer features thus being more successful. Interestingly, a moderate negative linear relationship also exists between tricky words and number of *st*- feature games played (r_s = -.430, p = .001), showing that participants who were more likely to save tricky words were less likely to repeat this feature.

Participants' gender also yielded weak relationships with WA gains (r = .268, p = .011), and NWA gains (r_s = .255, p = .028). Combined with the above relationships, these findings point to gender contributing to differences in game activity and gains, with females reading more books, saving more tricky words, and experiences greater word and non-word accuracy gains.

6. Discussion

6.1. iRead Gameplay and Gains

As an exploratory study, no specific hypothesis was put forward regarding the effect of gameplay on accuracy and fluency gains. Unlike Vanbecelaere et al (2020), Kartal and Terziyan, (2016), and van de Ven (2017), this study did not include a condition in which participants were not exposed to the game, but sought to tie accuracy and fluency gains to use of and performance in the game. Because the above mentioned studies have demonstrated that serious games can effectively contribute to short and long-term text, word and pseudo-word reading fluency and phonological awareness, there were again some general expectations about game use and how this may have affected overall gains.

With the first research question, for example, one may have expected to find that a lower number of games per feature led to greater gains due to higher proficiency leading to more rapid reading development. Conversely, one may have expected to find that a larger number of games per feature led to greater gains due to a feature being repeated until learned. Lastly, because of the high degree of variety in the number of features played, one may have expected to find greater gains in participants who played more games and features in total, indicating exposure to

a greater number of linguistic features. In line with Vanbecelaere et al (2020) and van de Ven (2017), our participants did experience gains in fluency with a mean increase of 14.92 words per minute between pre-test and posttest. However, no significant relationships were found between gains and game use. This finding may be interpreted in a number of ways.

One interpretation is that gains must be explained by external factors: exposure to L2 media, tutorial classes, etc. This conclusion would point to the ineffectiveness of the iRead application. However, a weak relationship between the WA-T1 measure and number of features played indicates that gameplay was influenced somewhat by initial proficiency level, and is an indicator of the adaptive nature of iRead. Should this be the case, it would point to the effectiveness of adaptivity. It shows that participants who initially had a higher level did not necessarily learn more from the games, but that less-proficient participants may have had the opportunity to play more games per feature and still show overall progress. Another interpretation is that the measures of gameplay were not appropriate to compare to gains. For example, total number of games and features were considered, but overall time spent playing was not. Also not taken into consideration was total number of log-ins, or mean time needed to complete games. This leaves the possibility that gains were impacted by gameplay and require further investigation in order to be identified.

Similar results were encountered when answering the second research question, which looked for a relationship between game performance and gains. This study operationalized game performance as number of successes divided by the square root of total games played. Also, number of games per feature was taken as a measure of game performance because failures resulted in the same feature being repeated. One may have expected to find a relationship between overall gains and success rate, where a positive correlation would indicate more successful participants exhibiting greater gains and a negative relationship indicating the opposite.

Because no relationships were established, we then looked at participant performance in 5 game-features that corresponded to items on the WA measure. Of the five features, only those pertaining to *st-* words yielded a moderate relationship between gains and number of times played. Though moderate, this finding was important because it is evidence of accuracy gains being tied to game performance; participants who experienced greater gains played this feature

more times than those who did not. This finding is in line with Kartal and Terziyan's (2016) findings in which participants who used a serious game focusing on phonological awareness were able to outperform participants who did not. (That said, Kartal and Terziyan did not focus on game performance.) Furthermore, this finding is especially welcome considering that the accuracy measure was not designed to investigate iRead feature use, and the number of WA items corresponding with iRead features was quite small.

There are a number of findings that may have been expected. One may have expected, for example, to find that lower-proficiency learners that played a larger number of total features or total games should have experienced greater gains. Conversely, and in accordance with the simple view, one may have expected that higher proficiency participants would have played a higher number of total games due to greater reading fluency, thus experiencing greater gains (Cain, 2015; Hoover and Gough, 1990). The data appears to support a more complex interpretation. A weak negative relationship was shown to exist between CEFR level and games per feature. This demonstrates that higher proficiency participants as defined by the CEFR rubric tended to require less games to successfully complete a feature, which is to be expected. A weak positive relationship was also found between reading proficiency and number of features played. This shows that participants that were able to read more quickly and fluently were able to unlock a greater number of features, in line with the simple view of reading (Hoover & Gough, 1990).

These findings are further supported by the negative relationships found pertaining to reading proficiency, CEFR rating and number of *st*-feature games, and the positive relationship between number of *st*.-feature games and *st*- gains. Considered together, these findings illustrate iRead functioning as it was designed to: allowing more proficient users to advance while forcing users who do not demonstrate mastery of a concept to revisit unlearned features. Thus, less proficient users were able to experience greater accuracy gains while more proficient users may have already demonstrated a mastery of the st-feature on the pretest thereby experience lesser gains.

Reading proficiency also moderately correlated with T1 and T2 measures of WA and WPM, but not gains. This not only reinforces that accuracy and fluency are components of word recognition (Cain, 2015; Hoover & Gough, 1990) but also supports the validity of the Cambridge

PET exam and may also be interpreted in a number of ways. Again, one may simply conclude that the null-effects in overall gains are evidence of iRead's ineffectiveness. However, as established above, this does not appear to be the case.

Relationships between CEFR level, reading proficiency, *st*-feature gains, and gameplay indicate that proficiency does indeed have an effect on gameplay. Also, the lack of data for error rate per feature means that the existence of yet to be discovered relationships cannot be discounted. Further consideration about the relationships between individual features, error rates, and gains need to be investigated. Moderate and strong relationships between pre-test and posttests and proficiency measures indicate that these measures are valid and effective at measuring what they are designed to measure. However, the lack of substantial evidence for gains relating to gameplay and gains relating to proficiency indicate that the source of these gains has yet to be explained. A pre-test and post-test measure designed around features that are anticipated to be encountered during iRead use may aid in further explaining the source of these gains. Additionally, lack of stronger relationships between gains and game use may be the result of a failure to take into account secondary mechanisms involved in L2 reading acquisition that do not exist in the L1 such as those mentioned by Grabe (2009), Nassaji (2011), Bernhardt and Kamil (2006) and Brevik et al (2016).

6.2. The Role of Adaptivity

For this study, the number of games played per feature was taken as a measure of game play but also speaks to adaptivity. As previously touched upon, adaptivity is something that sets serious games apart from traditional classroom settings, and there was some expectation that game play would at least partially contribute to WA and WPM gains by responding to the learning pace of individual participants. Though difficult to operationalize, there were a few relationships within the data that shed light onto its existence and function. Returning to the *st*-feature, a moderate negative linear relationship (r = -.307, p = .019) exists between T1 and number of games. This illustrates that participants who demonstrated mastery of *st*- items on the pre-test encountered the *st*- feature fewer times. iRead's adaptivity, is thus a double edged sword when attempting to explain WA and WPM gains with game use. Lower proficiency participants were able to revisit unlearned concepts until they were learned, while more proficient students

naturally progressed. This in turn created a situation in which 'a rising tide lifted all boats,' making it impossible to tie gameplay with overall gains due to ceiling effects.

In L2 classrooms, a teacher will normally teach to the abilities of middle to lower-proficiency students; thus slowing the developmental speed of the most proficient learners. One may expect the data to support this by showing lower and moderate proficiency participants to experience greater gains that higher proficiency students. Moderate positive linear relationships between gains in WA, NWA, and WPM show that this was not the case among our participants, and shows that the iRead algorithm was adapting to individual proficiency as outlined by Deligiannis et al (2019).

Adaptivity offers an avenue of future investigation in which a narrow focus may yield stronger results. For example, exploring the individual correct and incorrect responses within features may offer a glimpse into relationships between gains and how the algorithm responds to individual learner responses. With a 'success' threshold of 70%, is there a relationship between gains and participants who passed with 90% or 100% as opposed to those who passed with 70%? The effectiveness of differing types of corrective feedback is another component of iRead's adaptivity that requires future study. As with any technology that is still in its infancy, many conclusions have yet to be drawn. Considered within the context of this study, iRead's adaptivite algorithm may not have received the emphasis it deserved and may ultimately have been responsible for a lack of strong linear relationships within the data.

6.3. Books and Tricky words

Along with regular gameplay, participants also used iRead's Amigo e-reader feature to varying degrees ranging from no books read to 136 books read, with an average of 56.13 books and a standard deviation of 32.61 books during the four months of use. While reading, students also highlighted and saved unknown 'tricky words' that they encountered ranging from saving no tricky words to 130 words with the mean being 12.82 and a standard deviation of 25.37 words saved. Though not originally part of our design, these data offered another avenue of exploration with which we could better understand and explain accuracy and fluency gains within the context of iRead use. As shown in section 5.4., numerous weak and moderate relationships were identified between number of books, number of tricky words, and other datasets. Weak negative relationships between tricky words and gameplay features indicate that some participants were

more inclined to read stories while others played more games. This may be somewhat expected. Unexpectedly, however, were the weak positive relationships between number of tricky words and WA, NW, and WPM gains. Without straying too far outside of the scope of this study, this may indicate that individual differences or motivation may have also contributed to accuracy and fluency gains. Coupled with weak positive relationships between gender and WA and NWA gains, these data point to other unexplored factors bearing partial responsibility for L2 reading development among the participants.

7. Strengths and Limitations

As part of an ongoing investigation into the effectiveness of iRead, our study benefitted from an abundance of internal and external data with which we could design a study that simultaneously explored L2 reading development and game use. Certain findings of this study fall in line with the predictions put forward by the simple view of reading, indicating a strong design. As an exploration, the large amount of data allowed us to look for relationships between numerous factors. At times, this large amount of data also proved to be a blessing and a curse.

A major issue encountered stemmed from the WA measure as it relates to feature use. The WA measure we used was not related to iRead and did not consider the features that may have been encountered during gameplay. Had we used an accuracy measure that was designed around features that were expected to be encountered, the likelihood of identifying relationships between individual features and WA gains would have been greatly increased. Should a replication study be conducted, this is weakness needs to be addressed by developing a novel WA measure that incorporates words and linguistic features that are frequently seen in iRead.

Additionally, the way game performance was operationalized may have been poor because of its binary design. Simply analyzing game performance as a success or a failure leaves an area of game performance unexplored. A measure of error rates per feature would have been a stronger reflection of individual game performance, however this data was not available to researchers at the time of analysis. Further complicating matters is that specific games contain differing numbers of items, making them difficult to generalize. Using a rate of errors per feature would have offered a more detailed and thorough glimpse into gameplay and allowed us to better answer our second research question.

This study would have also greatly benefitted from a control group. A control group would have allowed us to definitively attribute accuracy and fluency gains to iRead use and not an unaccounted-for external factor.

In summation, weak and moderate relationships between un-related measures speak to the construct validity of this study. However, should this study be replicated, the above mentioned weaknesses must be considered and addressed.

8. Conclusion

The findings of this study not only support Cain's (2015) model of the simple view, but also suggest that the mechanisms involved in predicting L1 reading proficiency are at least partially mirrored in the L2. For example, relationships between accuracy and fluency pre-tests and gains and reading proficiency support that fluency and accuracy form a measure of word recognition. Further relationships between fluency and accuracy and listening indicate that the two are related. Finally, relationships between vocabulary, listening and reading proficiency as well as accuracy and fluency support the inclusion of vocabulary in the simple view's model.

Further research into the role of vocabulary in L2 reading acquisition may provide clarification and allow for an L2-specific simple view model to be proposed. Studies on the effectiveness of iRead as a L2 learning tool are ongoing. As evidenced by other studies which investigate the effectiveness of serious games, the weak effects found in answering our research questions should not be taken as a condemnation of serious games as a means of L2 instruction. The role of adaptivity and the appropriateness of the features selected by the application need to be further investigated. Furthermore, longitudinal studies comparing the relative effectiveness of serious games and traditional teacher-led classes may provide insights into the value of personalized and adaptive games as a mode of further language instruction and specifically L2 reading development.

Word Count: 12,353.

9. References

Adlof, S.M., Catts, H.W., & Little, T.D. (2006). Should the simple view of reading include a fluency component? *Reading and Writing*, *19*(9), 933–958. doi:10.1007/s11145-006-9024-z

Alderson, J.C.,Haapakangas, E.L.,Huhta, A.,Nieminen, L., & Ullakonoja, R. (2014). *The diagnosis of reading in a second or foreign language.* Routledge.

Bernhardt, E. & Kamil, M. (2006). *Encyclopedia of Language & Linguistics*. Elsevier Ltd.

Brevik, L.M., Olsen, R.V., & Hellekjär, G.L. (2016). The complexity of second language reading: Investigating the L1-L2 relationship. *Reading in a Foreign Language, 28*(2) 161-182.

Cain, K., Oakhill, J., Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96*(1), 31-42.

Cain, K. (2015). Learning to read: Should we keep things simple? *Reading Research Quarterly, 50*(2), 151-169.

Connoly, T., Boyle, E., MacArthur, E., Hainey, T., & Boyle, J. (2012). A systemic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*(2), 661-686.

Deligiannis, N., Panagiotopoulos, D., Patsilinakos, P., Raftopoulou, C.N., & Symvonis, A. (2019). Interactive and Personalized Activity eBooks for Learning to Read: The iRead Case. iTextbooks@AIED.

Foorman, B., Petscher, Y., Schatschneider, C. (2015). *Florida Assessments for Instruction in Reading, Aligned to the Language Arts Florida Standards, FAIR-FS, Grades 3 through 12: Technical Manual*. Tallahassee: Florida Center for Reading Research.

Fuchs, L., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256.

Gomez, S., Zervas, P., Sampson, D. G., & Fabregat, R. (2014). Context-aware adaptive and personalized mobile learning delivery supported by UoLmP. *Journal of King Saud University – Computer and Information Sciences*, *26*(1), 47-61

Gough, P., Hoover, W.A., & Peterson, C.L. (1996). Some observations on a simple view of reading. In C.Conoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and interventions*. (pp. 1-13). Erlbaum

Gough, P.B., & Tunmer, W.E. (1986). Decoding, reading and reading disability. *Remedial and Special Education, 7*(1), 6–10. doi:10.1177/074193258600700104

Grabe, W. (2009). *Reading in a second language: moving from theory to practice*. Cambridge University Press.

Holmes, W., Anastopoulou, S., Schaumburg, H., & Mavrikis, M. (2018). *Technology-enhanced personalised learning: Untangling the evidence.* Robert Bosch Stiftung GmbH.

Harris, T. L., Hodges, R. E. (1995). *The literacy dictionary: the vocabulary of reading and writing.* International Reading Association

Hoover, W. A., Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, *2*, 127-160. https://doi.org/10.1007/BF00401799

Johnston, T.C.,& Kirby, J.R. (2006). The contribution of naming speed to the simple view of reading. *Reading and Writing*, *19*, 339-361. https://doi.org/10.1007/s11145-005-4644-2'

Kartal, G., & Terziyan, T. (2016). Development and evaluation of game-like phonological awareness software for kindergarteners: JerenAli. *Journal of Educational Computing Research, 53*(4), 519–539.

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning, 57*, 1-44. http://dx.doi.org/10.1111/0023-8333.101997010-i1

Kyle, F., Kujala, J., Richardson, U., Lyytinen, H., & Goswami, U. (2013). Assessing the effectiveness of two theoretically motivated computer assisted reading interventions in the United Kingdom: GG rime and GG phoneme. *Reading Research Quarterly*, *48*(1), 61–76.

Martin, E. & Carro M. (2009). Supporting the development of mobile adaptive learning environments: A case study. *IEEE Transactions on Learning Technologies, 2*(1). 23-36. doi: 10.1109/TLT.2008.24

Michael, D. & Chen, S. (2006). *Serious games: Games that educate, train, and inform.* Muska & Lipman/Premier – Trade.

Mitchell, A.M., Brady, S.A. (2013). The effect of vocabulary knowledge on novel word identification. *Annals of Dyslexia, 63*, 201–216. https://doi.org/10.1007/s11881-013-0080-1

Nassaji, H. (2011). Issues in second-language reading: Implications for acquisition and instruction. *Reading Research Quarterly*, *46*(2), 173-184.

Nation, K., & Snowling, M.J. (1998). Individual differences in contextual facilitation: Evidence from dyslexia and poor reading comprehension. *Child Development*, *69*(4), 996-1011. doi:10.1111/j.1467-8624.1998.tb06157.x

Nation, K., & Snowling, M.J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading, 27*(4), 342– 356. doi:10.1111/j.1467-9817.2004.00238.x

Nation, P. & Anthony, L. (2016). Measuring vocabulary size. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning,* (3rd vol., pp. 355-368). Routledge.

Ouellette, G.P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. Journal of Educational Psychology, *98*(3), 554–566. doi:10.1037/0022-0663.98.3.554

Ouellette, G., Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing, 23*, 189–208. https://doi.org/10.1007/s11145-008-9159-1

Pattemore, M., Gilabert, R., & Serra, J. (2019). Elaborative Feedback in L2 Reading Games. Workshop paper presentation at GamiLearn 2019 conference. October 22, 2019. Barcelona, Spain.

Perfetti, C.A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. doi:10.1080/10888430701530730

Protopapas, A., Simos, P.G., Sideridis, G.D., & Mouzaki, A. (2012). The components of the simple view of reading: A confirmatory factor analysis. *Reading Psychology*, *33*(3), 217–240. doi:10.1080/02702711.2010.507626

Stanovich, K.E., & West, R.F. (1979). Mechanisms of sentence context effects in reading: Automatic activation and conscious attention. *Memory & Cognition*, *7*(2), 77–85. doi:10.3758/BF03197588

Susi, T., Johannesson, M., & Backlund, P. (2007). Serious games – An overview. *IKI Technical Report.* School of Humanities and Informatics, University of Skövde, Sweden.

Vanbecelaere, S., Van den Berghe, K., Cornillie, F., Sasanguie, D., Reynvoet, B., Depaepe, F. (2020). The effects of two digital educational games on cognitive and non-cognitive math and reading outcomes. *Computers & Education, 143*

van de Ven, M., de Leeuw, L., van Weerdenburg, M., & Steenbeek-Planting, E. G. (2017). Early reading intervention by means of a multicomponent reading game. *Journal of Computer Assisted Learning, 33*(4), 320–333.

Wu, S., Chang, A., Chang, M., Liu, T.C., Heh, J. S., (2008). Identifying personalized context-aware knowledge structure for individual user in ubiquitous learning environment. In: *Proc. 5th International Conference on Wireless, Mobile and Ubiquitous Technologies in Education* (pp. 95–99).

9. Appendixes

Appendix A: Word reading accuracy word-list:

And	We	She	Pet	Help	Shop	Have
Old	When	Like	They	Two	Ball	Girl
Story	Jumped	Saw	Such	Think	Because	Slow
Joy	Came	Each	First	Splash	Train	Always
Stood	Pencil	Might	Those	Toe	Weather	Trick
Knee	Dreaming	Fruit	Between	Painting	Wrong	Understand
Pole	Length	Cage	Proud	Nervous	Straight	Company
Emergency	Crystal	Island	Melody	Achieve	Anniversary	Radiator
Creation	Excitement	Generally	Leisure	Guilt	Queue	Organic
Knowledge	Probability	Crisis	Authority	Gnaw	Enthusiasm	Distinguish
Nuisance	Society	Quadruple	Pragmatic	Contagious	Signify	Ridicule
Misconception	Debt	Deceit	Pronunciation	Delinquent	Abominable	Sieve

Vivacious	Endeavor	Melancholy	Mediate	Placebo	Euphemism

Appendix B: Non-word word list

Ip	Ga	Ko	Ta	Om	Ig
Ni	Pim	Wum	Lat	Baf	Din
Nup	Fet	Bave	Pate	Herm	Dess
Chur	Knap	Tive	Barp	Stip	Plin
Frip	Poth	Vasp	Meest	Shlee	Guddy
Skree	Felly	Clirt	Sline	Dreff	Prain
Zint	Bloot	Trisk	Kelm	Strone	Lunaf
Cratty	Trober	Depate	Glant	Sploosh	Dreker
Ritlun	Hedfert	Bremick	Nifpate	Brinbert	Clabom
Drepnort	Shrattec	Plofent	Smucrit	Pelnador	Fornalask
Fermabalt	Crenidmoke	Emulbatate	Strotalanted	Prilingdorfent	Chunfendilt

Appendix C: Oral reading fluency passage: How to Make Play Dough

Script: I would like you to read out loud for me. I will use my stopwatch to tell me when I want you to stop reading. Please do your best reading because I will ask you a question about what you read. Do you understand what we will be doing? This story is called How to Make Play Dough. Begin here. Ready? (Point to the first word of the text. Start the stopwatch when the student reads the first word.)

How to Make Play Dough

It was raining today. I couldn't play outside. I couldn't find my old red truck. I was feeling kind of blue. My mom could tell I was bored. She asked me if I wanted to make play dough. Boy was I excited! It was so much fun!

If you want to make your own play dough, you need to get a grown-up to help. First, get a large bowl. Mix one cup of flour, one cup warm water, two teaspoons cream of tartar, one teaspoon oil and ¼ cup salt. Then, use your hands to kneed (mix) it together. Add food coloring if you want to make colored dough. Now ask the grown up to help you put the dough in a pot. Place the pot on the stove. Stir the dough over medium heat until the lumps are gone. Ask the grown-up to remove it from the hot pan. Let the dough cool. Then kneed it some more until smooth. Not it is ready to play!

I made five batches of dough. I had red, blue, green, yellow, and black. I made animals and bugs with my play dough. My mom made eight big cookies. They looked good enough to eat! Then she made a cup. I had a pretend drink with my pretend cookies! Later, I helped my mom wash the dishes we used. We cleaned the table where we played. We put all the dough in plastic bags so it would stay soft. I decided to keep one of my creations. I left the very long and pretty caterpillar out to dry out. I put it on the shelf in my room. Now I will always remember the fun I had on this rainy day!

Appendix D: Descriptives

Measure:	*N=*	*Min.*	*Max.*	*M=*	*SD=*

WA-T1	72	20	69	45.70	10.65
WA-T2	72	25	78	51.00	10.73
WA Gains	72	-13	26	5.89	9.10
NWA-T1	72	26	60	45.42	9.60
NWA-T2	72	30	64	47.06	7.47
NWA Gains	72	-16	18	1.51	7.44
WPM-T1	72	40	144	89.71	21.80
WPM T2	72	57	148	103.97	20.18
WPM Gains	72	-15	42	14.92	13.63
Numb. Games	72	28	227	120.74	44.76
Numb. Features	72	21	72	47.82	11.86
Games/Feature	72	1.21	3.55	2.44	.59
Numb. Success	72	20	222	110.39	43.86
Success/Root	72	4.33	14.73	9.79	2.32
Success/Game	72	.782	.987	.903	.053
Listening	72	2	19	9.17	3.2
Reading	72	5	32	14.64	4.82
Vocabulary	72	28	89	52.59	9.38
CEFR	72	1	3	1.47	.556
Books	72	0	136	56.13	32.61
Tricky Words	72	0	130	12.82	25.37
Kn-T1	46	0	2	.67	.668
Kn-T2	46	0	2	1.13	.718
Kn-Games	46	0	9	2.48	1.88
Kn-Gains	46	-2	2	.46	.808
U-T1	46	0	1	.70	.465
U-T2	46	0	1	.65	.482
U-Games	46	0	10	2.04	1.885
U-Gains	46	-1	1	-.04	.556
Spl-T1	46	0	1	.89	.315
Spl-T2	46	0	1	.56	.482
Spl-Games	46	0	9	1.87	1.655
Spl-Gains	46	-1	1	.04	.419
St-T1	46	1	3	2.46	.690
St-T2	46	0	3	2.39	.774
St-Games	46	0	12	2.28	2.509
St-Gains	46	-2	3	-.07	.800
-o_e-T1	46	0	3	1.50	.913
-o_e-T2	46	0	3	2.46	.690
-o_e-Games	46	0	8	2.50	2.052
-o_e-Gains	46	-2	3	.09	1.029

Note: WA-T1 = word accuracy at pre-test; WA-T2 = word accuracy at posttest; WA Gains = word accuracy gains; NWA-T1 = non-word accuracy at pre-test; NWA-T2 = non-word accuracy at posttest; NWA Gains = non-word accuracy gains; WPM-T1 = fluency at pre-test; WPM-T2 = fluency at posttest; WPM Gains = fluency gains; Numb. Games = number of games played; Numb. Features = number of features played; Games/Feature = number of games played per feature; Numb. Success = number of games which ended in 'success'; Success/Root = number of successes divided by the square root of number of games played; Success/Game = number of successes divided by the number of games played; Listening = score from listening portion of PET; Reading = score from reading portion

of PET; Vocabulary = score from vocabulary portion of PVST; CEFR = common European framework of language rating where A1 = 1, A2 = 2, B1 = 3, etc…; Books = number of Amigo books read; Tricky Words = number of tricky words saved in Amigo; Kn-T1 = correct kn- items at pre-test; Kn-T2 = correct kn-items at posttest; Kn-games = number of features played containing kn-; Kn-Gains = gains in kn- between pre-test and posttest; U-T1 = correct u- items at pre-test; U-T2 = correct u-items at posttest; U-games = number of features played containing u-; U-Gains = gains in U- between pre-test and posttest; Spl-T1 = correct spl- items at pre-test; Spl-T2 = correct spl- items at posttest; Spl-games = number of features played containing spl-; Spl-Gains = gains in spl- between pre-test and posttest; St-T1 = correct st- items at pre-test; St-T2 = correct st-items at posttest; St-games = number of features played containing st-; St-Gains = gains in st- between pre-test and posttest; O_e-T1 = correct o_e- items at pre-test; O_e-T2 = correct o_e-items at posttest; O_e-games = number of features played containing o_e-; O_e-Gains = gains in o_e- between pre-test and posttest;

APPENDIX E: Preliminary analysis.

In order to test the reliability of WA, NWA, and WPM, two tailed correlations were run between pre-test T1 and T2 data. Measures for WA showed a moderate positive linear relationship between T1 and T2, (r_s = .509, p < .001). A moderate positive linear relationship was also shown to exist between T1 and T2 NWA measures, (r_s = .621, p < .001). Finally, a strong positive linear relationship is shown to exist between T1 and T2 WPM measures, (r_s = .778, p < .001). These relationships show that the tests were effective indicators of individual participants' word and non-word reading accuracy and fluency over time.

Moderate positive linear relationships also exist between WA, NWA, and WPM measures. For example, WA-T1 data shows moderate positive linear relationships with NWA-T1 (r_s = .555, p < .001), NWA-T2 (r_s = .357, p = .002), WPM-T1 (r_s = .682, p < .001) and WPM-T2 (r_s = .483, p < .001). Similarly, WA-T2 data shows moderate positive linear relationships with NWA-T1 (r_s = .413, p < .001), NWA-T2 (r_s = .592, p < .001), WPM-T1 (r_s = .425, p < .001), and WPM-T2 (r_s = .531, p < .001). These moderate positive linear relationships indicate that the WA, NWA and WPM measures are well-designed, evenly weighted and generalizable.

As for proficiency measures, tests for relationships between listening, reading, vocabulary, and CEFR-level proficiency measures were conducted using correlations. Listening proficiency showed moderate positive linear relationships with reading proficiency (r_s = .392, p < .001), vocabulary (r_s = .314, p = .004) and CEFR level (r_s = .322, p = .003). Reading proficiency also showed moderate positive linear relationships with vocabulary (r_s = .432, p < .001) and a strong positive linear relationship with CEFR level (r_s = .802, p < .001). A moderate positive linear relationship was also found between vocabulary and CEFR level (r_s = .446, p < .001). These relationships speak to the validity of the PET and PVST as well as the CEFR language proficiency rubric as generalizable indicators of overall proficiency.

Tests were also performed measuring relationships between the T1 and T2 scores for WA, NWA, and WPM and the proficiency measures obtained from the PET and PVST. Numerous weak and moderate relationships were shown to exist. Listening proficiency scores correlated weakly with WA-T1 (r = .209, p = .013) and moderately with WPM-T1 (r = .486, p < .001), and WPM-T2 (r = .483, p < .001). Similarly, there was a weak positive relationship between reading proficiency and WA-T1 (r = .269, p = .022), and moderate positive relationships with WPM-T1 (r = .387, p = .001), and WPM-T2 (r = .440, p < .001). Moderate positive relationships were also found to exist between vocabulary proficiency and WA-T1 (r = .393, p = .001), WA-T2 (r = .334, p = .004), WPM-T1 (r = .503, p < .001), and WPM-T2 (r =

.392, p = .001). Though somewhat to be expected, these relationships not only speak to the validity of the individual measures involved, but also support the simple view's theoretical framework, which will be discussed at length in the discussion section.