Play games or study? Computer games in eBooks to learn English vocabulary

Glenn Gordon Smith [a,*], Mimi Li [a], Jack Drobisz [a], Ho-Ryong Park [b], Deoksoon Kim [a], Stanley Dana Smith [c]

[a] University of South Florida, 4202 E. Fowler Ave., Tampa, FL 33620-5650, USA
[b] Murray State University, 100 Faculty Hall, Murray, KY 42071, USA
[c] Hawaii Pacific University, 1166 Fort Street Mall, Honolulu, HI 96813, USA

ARTICLE INFO

ABSTRACT

This study investigated how Chinese undergraduate college students studying English as a foreign language learned new vocabulary with inference-based computer games embedded in eBooks. The investigators specifically examined (a) the effectiveness of computer games (using inferencing) in eBooks, compared with hardcopy booklets for vocabulary retention, and (b) the relationship between students' performance on computer games and performance on a vocabulary test. A database recorded students' game playing behaviors in the log file. Students were pre- and post-tested on new vocabulary words with the Vocabulary Knowledge Scale. Participants learned significantly more vocabulary ($p < .0005$) in the computer game condition (web-based text and computer games) than in the control condition (their usual study method, hardcopy text, lists of words and multiple-choice questions). Students' scores in the games correlated significantly with their vocabulary post-test scores ($r = .515$, $p < .01$).

1. Introduction

Learning English, for Chinese undergraduate students, is both a global imperative, and an enormous challenge. It is a global imperative because English is the Lingua Franca, the dominant language of the world today, the international language of business, science, and culture (Smith, 2005). China, as a rising economic superstar, needs a workforce fluent in this international language. It is an enormous challenge because Chinese and English are vastly different languages from different typologies (Haynes, 1990; Wu, Lowyck, Sercu, & Elen, 2013). Asian immigrants to the United States, including those who first language was Mandarin, learned English less well than immigrants from European countries (Jia, Aaronson, & Wu, 2002). Chinese (Mandarin) frequently uses intonation, i.e., changes in pitch to differentiate vocabulary to convey semantic meaning, in contrast to English, which relies more on morphology and word sequence. Chinese writing is primarily logographic (each symbol represents a word), while English writing is primarily alphabetic (each symbol, or combinations of symbols, represent phonemes, but with highly inconsistent rules). Beyond these differences, learning English is difficult because English has a larger vocabulary than any other language (Sewell, 2008). Furthermore, English, reflecting various invasions, a colonial history, a willingness to take on all linguistic comers, draws words from a bewildering number of other languages (Sewell, 2008). Inconsistent spelling rules reflect this spotted etymology.

The challenge of learning English for many Asian university students, including those whose native language is Chinese, Chinese-related languages or Korean is supported by cross-linguistic second language acquisition studies (Gan, Humphreys, & Hamp-Lyons, 2004; Gu & Johnson, 1996; Wu et al., 2013). A study by Flege, Jeni-Komshian, and Liu (1999) suggest that Asians such as Koreans learning English as teenagers or adults do not master English phonology or grammar as well as those who learn earlier. In contrast, research conducted on the grammatical competence of bilinguals whose first language is Spanish or Dutch–languages that are more similar to English–does not show any consistent relationship between age of arrival in the US and mastery of English grammar (Birdsong, & Molis, 2001).

Aside from native differences in L1 and L2 languages, there are other factors that make learning English challenging for Chinese university students. Many Chinese university students embrace tedious study practices, such as rote memorization of lists of words without using the words in context (Gan et al., 2004, p 236). One of the authors of the current paper notes that the English language courses at the university where she taught in China, Sichuan Normal University (SNU), prescribed rote memorization of lists of words (along with a text passage containing the words) as standard practice, and that the students complained about the tediousness of such practices. The current study seeks to compare the traditional study practices of English vocabulary in one Chinese university, with a game-based approach.

A number of studies have investigated the potential *incidental* benefits of commercial computer games on L2 vocabulary (e.g., Thorne, Fischer, & Lu, 2012). More relevant to the current study, Cobb & Horst (2011) investigated *intentional* L2 vocabulary learning with a commercial computer game, Word Coach, explicitly designed for intentional L2 vocabulary learning. In a within subjects quasi-experiment using whole classes, both classes played Word Coach for two months (but during different time periods), children (11–12 years old) learned significantly more new English vocabulary after playing Word Coach for two months, versus without Word Coach.

However, it appears that no one has investigated an L2 intentional game-play vocabulary learning intervention designed for a specific course and formal learning situation. The current study investigated a specific L2 vocabulary computer game intervention designed for Sichuan Normal University English courses. The current study focused on how students *study* new English vocabulary outside the classroom for this specific English course and the materials they use for this studying.

The investigators acknowledge that educational game play and traditional study methods are made up of many different factors and components. For instance games provide built-in incentives. However, the goal of the current study was to provide an initial assessment as to whether a game-based approach may provide a more effective way for Chinese undergraduates to learn English vocabulary. That is, initial evidence supporting a game-like approach may provide impetus for other studies to isolate and test the different components of the game play approach. Accordingly, the current study investigated a promising online game play solution in its entirety, with respect to how this new solution might be an improvement over traditional methods of studying English vocabulary. The current study, in terms of method, is classified as a *design experiment* (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Reinking & Bradley, 2008). The design experiment, a relatively new method in educational research, borrows Engineering methods and applies them to educational research. An engineer, confronted with a problem, comes up with an idea, designs a solution, builds a whole prototype, and tests it rigorously to see to what extent it solves the problem, and to what extent it improves on existing solutions. The designing, building and testing of the solution often sheds light on theoretical ideas. The design experiment does not seek to isolate one variable, but rather implements and tests a whole solution, and evaluates it along multiple dimensions. Such an approach is both pragmatic and theoretical, as it suggests a possible new educational approach as well as possible theoretical factors to investigate in isolation via other research methods.

Design experiments focus on the "learning ecology" of a specific learning situation (Cobb et al., 2003). There are five crosscutting features of design experiments (Cobb et al., 2003): First, the purpose is to develop theories about both the learning process, and the means to support that learning for that specific situation. The second feature is the interventionist nature of the research. Thirdly, design experiments test conjectures about the learning processes in a particular situation, but also potentially generate new conjectures to test.

Fourthly, design experiments are often iterative, as conjectures are generated and refuted.

Fifthly, theories developed in the design experiment are humble, domain-specific, but also have implications from the design experiment (theory influences the design and the outcome of the design influences the theory).

As such a design experiment often involves a number of steps: (1) "initial design of an intervention based on current theoretical understanding, with an explicit underlying causal explanation of the causal effect" (Gorard, Roberts, & Taylor, 2004), (2) formative evaluation of the intervention using qualitative techniques, and (3) feasibility of the intervention, with measurement and in-depth feedback. The current paper describes one such study of game-play study techniques for L2 English vocabulary learning.

This study investigated how computer games, using inferencing, can help Chinese college students studying English as a foreign language (EFL) learn new vocabulary. Acquiring new vocabularies is always emphasized for EFL students because it provides a foundation to build on in mastering English. The researchers in this study were interested in bringing innovation to EFL college students' vocabulary learning by embedding it in a reading context within a game context. The computer game design was informed by the constructs of deep processing and inferences, which have theoretical implications in second-language vocabulary learning (Ellis, 1995). This study specifically compared the effectiveness of web-based computer inferencing games with that of EFL learners' typical vocabulary-learning practices using hardcopy materials. The investigators also examined the relationship between game performance and vocabulary learning.

1.1. Vocabulary learning in the Chinese EFL context

Vocabulary learning is an important part of college English courses in China. Specific vocabulary sizes are stipulated for the basic, intermediate, and higher requirements in the college English curriculum. At the basic level, students need to acquire a total of 4795 words and 700 phrases, including 2000 active words, which they must not only comprehend, but also use fluently in speaking and writing. At the intermediate level, students must acquire a total of 6395 words and 1200 phrases, including 2200 active words. The higher level requires a total of 7675 words and 1870 phrases, including 2360 active words. Although college English courses attach great value to vocabulary instruction, acquiring a large English vocabulary is still one of the greatest challenges for the Chinese EFL college students. Students typically learn vocabulary by rote memorization of lexicon lists in their textbooks and from additional vocabulary books (Gan et al., 2004). According to one of the authors, who taught sections of English courses for Chinese students at Sichuan Normal University in China, most of her students regarded vocabulary learning as a chore. They complained that they invested much time in memorizing new words, but eventually remembered only a small portion of them. This situation calls for pedagogical innovation.

1.2. Computer assisted language learning (CALL) and computer games for vocabulary learning

With the emerging development of computer assisted language learning (CALL), many technology-incorporated vocabulary learning systems are designed to make vocabulary learning more interesting and more effective (e.g., Abraham, 2008; Başoğlu & Akdemir, 2010; Groot, 2000; Ma & Kelly, 2006; Oberg, 2011; Yun, 2011). Using multimedia in texts, including images and videos, has played an

important role in vocabulary acquisition (Chun & Plass, 1996; Kayaoglu, Dag Akbas, & Ozturk, 2011; Segler, Pain, & Sorace, 2002). As Ellis (1995) posited, when learners are provided both the reading context and the side-by-side definition on the screen in the CALL context, they can readily switch attention between the two, which greatly reduces their cognitive load (Sweller, 1994). In addition, hypertext messages, via mouse click, can provide an instant definition and explanation, giving the connotation of a word in context (Abraham, 2008; AbuSeileek, 2008; Yun, 2011).

Other studies explored gaming for a specific purpose in language learning, such as exploring learners' perceptions of corrective feedback in an immersive game for English pragmatics (Cornillie, Claebout, & Desmet, 2012), listening strategies (Roussel, 2011), and students' English as a foreign language (EFL) writing and speaking performance using a multimedia web annotation system (Hwang, Shadiev, & Huang, 2012). Two studies discussed virtual environments and online games to enhance vocabulary use (Bytheway, 2011; Rankin, Morrison, McNeal, Gooch, & Shute, 2009).

A few studies have investigated the adoption of commercial computer games, into university L2 courses. However, the generalizability of the findings from these studies is limited by small sample sizes. For example, Miller & Hegelheimer (2006) and Ranalli (2008) integrated the game, "The SIMS," along with associate supplemental activities, into university L2 courses with significant improvements in vocabulary learning, but with small sample sizes. However, Miller & Hegelheimer (2006) had only nine participants, and Ranalli (2008) only18. Furthermore, the significant improvements in vocabulary did not occur without supplementary materials involving the vocabulary (e.g., vocabulary lists and exercises, grammar descriptions and exercises, cultural notes, on-line dictionary, grammar explanation, and cultural notes). Hence, it is difficult to separate out the effects of the games versus the supplementary materials. In a third study, deHaan, Reed, & Kuwada (2010) studied 80 Japanese undergraduate computer science students in an English for Specific Purposes (computer science) course. Students were paired, pilot-copilot style, with one playing an English language music video game and the other observing. Both players and observers recalled vocabulary from the game, but the observers recalled significantly more than the players. Finally, Rankin, Gold, & Gooch (2006) conducted a pilot study with of five ESL learners of unspecified age playing the Massively Multiplayer Online Role-playing Game (MMORG) Ever Quest, with no comparison group. Rankin et al. claimed 40% in participant vocabulary increase, but the current authors find their sample size, methods and data analysis to be lacking in rigor. Most of the studies have repurposed recreational computer games for L2 vocabulary learning. Cobb & Horst (2011), who investigated Word Coach (designed for vocabulary learning), is the exception.

A body of research has investigated the benefits of multimedia, as opposed to computer games, for learning of L2 vocabulary, specifically how multimedia glosses (hyperlinked definitions) have been used in second language vocabulary acquisition in a CALL environment to enhance the acquisition of L2 vocabulary. Multimedia glosses for assisting learners in vocabulary acquisition consist of different modalities (textual, visual, and auditory) and modes (video, picture, and text) (Mohsen & Balakumar, 2011; Nation, 2001). The addition of multimedia (adding pictures, videos, sound, etc., to text) makes glosses more effective than they are with text alone (Gettys, Imhof, & Kautz, 2001; Martinez-Lage, 1997). Abraham's (2008) meta-analysis of studies of computer-mediated glosses indicated that glosses had an overall large positive effect on *incidental* L2 vocabulary learning. Lyman-Hager, Davis, Burnett, & Chennault (1993) reported that multimedia had a significant positive impact on vocabulary recall and retention. Overall, multimedia approaches to L2 vocabulary learning have been found to be beneficial.

In addition to multimedia glosses, computer-based environments can provide other affordances for L2 learners. In the second- or foreign-language learning context, positive learning effects of computer games and virtual worlds often stem from *cultural* and *lingual immersion*, collaboration with native speakers via immersive environments such as virtual worlds (Young et al., 2012; Zheng, Young, Wagner, & Brewer, 2009). In a meta-analysis of computer games used for education, broken down by discipline, Young et al. (2012) found few significant positive effects for the majority of disciplines, but significant positive learning effects for language learning. Young et al. (2012) speculated that the success of computer-game–based language learning might result from the social nature of language learning and the associated social nature of computer games. In addition, educational computer games can increase learners' motivation (Chen & Yang, 2013; Dickey, 2011). However, computer games may offer other nonsocial affordances for second-language learning: for instance, interaction with the game itself, the structure of the game-play, the challenge of the game, and the "*psychosocial moratorium*"—no bad consequences, e.g., no one really dies in war computer games (Gee, 2003)—which well deserve investigation.

1.3. Theoretical constructs motivating a design experiment in EFL vocabulary learning

The current researchers were interested in creating an intervention and design experiment (Cobb et al., 2003; Gorard et al., 2004; Reinking & Bradley, 2008) to improve Chinese undergraduate learning of L2 English vocabulary. As such they were interested in using the following theoretical constructs to create an intervention (change in practice) in how Chinese undergraduates study L2 English vocabulary.

In contrast to exploring immersive online gaming involving the interaction of multiple players, this study focuses on a single-player game, addressing some nonsocial effects of computer games, such as the design of the computer games, and discussing computer games' effects on students' vocabulary learning. Informed by the importance of *inferencing* in second-language reading and learning (Ellis, 1995; Mondria, 2003), as discussed in Section 1.5 below, the investigators designed an educational computer game for learning new vocabulary that would require EFL students to use new vocabulary words to make inferences about a text.

1.4. Deep processing and various conditions for vocabulary learning

The construct of *deep processing*, through inferencing, informed the investigators' design of the computer games in this study. Craik and Lockhart (1972) proposed a framework in which the effectiveness of encoding in long-term memory depends on how deeply the new information is processed. Shallow processing (e.g., oral rehearsal) does not lead to long-term retention, but deeper processing, whereby semantic associations are accessed and connected to other information in memory, does lead to long-term retention.

The levels of processing framework has been used for studies of memorization of new vocabulary items for both native and second language learners (Meyers, 2010; Shaunessy & Dinell, 1999). The forms of elaboration used in these studies include *sentence generation*. Sentence generation means simply that the participant creates a sentence using the new word. The depth of processing approach has been

extended to elaboration (Anderson & Reder,1979; Bradshaw & Anderson, 1982; Craik & Tulving, 1975; Mandler & Dorfman, 1994) and generative learning (Meyers, 2010). Generative learning (Wittrock, 1974) refers to the generation of semantic associations between the new information and items already in long-term memory, in the case of new vocabulary between the new word and existing knowledge or between the new word and other new words (Wittrock, 1990). Generative learning is sometimes misinterpreted as its generic meaning, i.e., learning through generating something such as writing a sentence with the new vocabulary. However, this more generic interpretation is more accurately described as productive learning, as opposed to receptive learning (Mondria & Wiersma, 2004). Examples are writing versus reading, or speaking versus hearing/comprehending.

The current investigators were interested in sentence generation as a form of elaboration, and generative learning. Meyers (2010) investigated sentence generation for second language vocabulary learning. He found larger beneficial effect sizes for second language vocabulary learning for generative tasks, which were more productive (e.g., writing), than for tasks that were more receptive (e.g., reading). Meyers (2010) also found that sentence writing tasks designed to maximize the generating of associations between new vocabulary words and background knowledge/experience resulted in significantly more vocabulary learning than sentence writing tasks designed to minimize generating such associations. However, since participants had free choice as to which sentences they wrote, Meyers (2010) also found it difficult to ensure that participants generated the kind of sentences specified in the experimental task. In the condition where the participants were instructed to write sentences with parameters likely to generate a lot of associations, many participants did not follow instructions. Later during data analysis, the experimenter had to check each sentence written by each participant, and weed out cases that did not fit with the experimental design.

The current authors sought to design a sentence generation task, with automated verification that avoided these problems through using an interactive game-like interface that constrained participant choices. In the above study (Meyers, 2010), one of the problems encountered was that since the sentence generation tasks were relatively open-ended, i.e., participants could choose what sentence to write, and that sentence might or might not conform to the conditions of generative learning (help the participant generate associations between the new word and their long term memory). The current authors sought to constrain the sentence writing by providing a context for generating associations, by providing a text passage containing the new vocabulary words, and by encouraging and constraining participants to create sentences that inferences from the text passages.

1.5. Inferences for vocabulary learning

Inferencing, i.e., determining the meaning of a new word from its context, is a key strategy for second- and foreign-language vocabulary learning (Ellis, 1995; Mondria, 2003). For L2 learners, finding the meaning of new words happens naturally when the encountering new words in context (either in speech or text). In an experimental context, L2 learners retain as much vocabulary through inferencing the meaning of words in context, as they do when provided with definitions (Mondria, 2003). However, learning vocabulary through inferencing alone requires more time than using definitions (Mondria, 2003).

Inferencing is not only an important process for L2 vocabulary, but is also a key component of text comprehension in general (L1 or L2). Readers must infer a great deal of information not explicit in the text to understand a text. Inferencing, considered the "heart of reading," is "the ability to use two or more pieces of information from a text to arrive at a third piece of information that is implicit" (Kispal, 2008).

A commonly accepted taxonomy of inferences includes 13 types of inferences from Graesser, Singer, & Trabasso (1994). For instance in one common type of inference, *causal antecedent*, the reader infers the cause of an event. An explanation of all 13 types of inferences goes beyond the scope of the current paper. The reader is directed to Graesser et al. (1994) for more information.

A simpler taxonomy of inferences, Trabasso & Magliano (1996) lists three types of inferences: (a) *backward* (explanatory) in which the reader uses the current sentence, to make inferences that form cohesion backwards to previous read sentences, (b) *forward* (predictive) inferences, in which the reader predicts what will happen next, making cohesion between the current sentence and upcoming sentences, and (c) *concurrent* (associative) inferences, where the readers makes connections between the current sentence and their own Long Term Memory.

These two types of inferencing (L2 inferencing of meanings of vocabulary and inferencing as part of the comprehension process) provide learning affordances that have great potential for game-based learning. The current study seeks to investigate the potential of using inferencing in game-based learning situation to provide L2 learners with a vocabulary learning activity that uses deeper cognition than what typically occurs in standard rote memorization. During deep processing, semantic associations are accessed and elaborated, which is likely to result in better vocabulary acquisition (Ellis, 1995).

1.5.1. Genesis of the design experiment

One of the PhD students in our research group, a Chinese national, was returning for six months (June–November) to the university in China (Sichuan Normal University) where she worked as an instructor before coming to a university in the South East of the United States to pursue her PhD. Prior to her trip, she discussed with her research group, in the U.S, the need for an improved study system for undergraduates at Sichuan Normal University for learning L2 English vocabulary. The research group (authors of the current paper) agreed that this would be an excellent opportunity for an intervention and design experiment on using game-play study activities and materials to improve studying of L2 new vocabulary.

During the six months while the Ph.D. student was in China, the research was conducted with two geographically distant parties, who communicated through email and internet-based teleconferencing: (a) the *main research group* in the United States (comprised of two professors, 10 masters students, and one Ph.D. student) who designed the intervention and developed instruments, and (b) the *onsite coordinator*, a Ph.D. student from the authors' research group, on-site at Sichuan Normal University, who helped design the intervention, interviewed instructors and students, found source materials for developing the intervention instruments which she emailed the group in the United states, conducted a pilot study and then a full intervention.

Our onsite coordinator emailed the development team in Florida materials that are sometimes used in the English courses at Sichuan Normal University, specifically 12 short text readings and the vocabulary lists that were associated with them.

We then considered the theoretical imperatives for our design process, the use of inferencing in L2 vocabulary learning and in text comprehension, as well as game play for motivation. Also, we wanted to conduct a design experiment to see if we could harness some of the

advantages of productive and generative tasks, such as sentence generation, that Meyers (2010) demonstrated to be more effective for learning L2 vocabulary than receptive tasks involving reading (true and false questions). However, we wanted to automate the sentence generation tasks to avoid the logistic difficulties encountered by Meyers (2010) who had to review the sentences generated by the participants to make sure the sentences written were indeed generative (like to create associations between the new vocabulary words and Long Term Memory (LTM)).

Based on these theoretical imperatives and using an educational game creation system designed to alternate computer games with text chapters (IMapBooks), we created eight text with game-segments, which we sent to our on-site coordinator in China.

1.6. Research questions

1) What is the effectiveness of the inferencing computer games compared with the hardcopy booklets for vocabulary retention? That is, how does vocabulary learning and new vocabulary retention compare under two conditions: a) reading text with new vocabulary embedded and playing online inferencing computer games (automated productive and generative task) and b) reading text in hardcopy with new vocabulary embedded and then using memorization lists or other conventional vocabulary-learning activities (more receptive task, and less generative task)?

The research question can be cast in another way: Can an automated, teacher or researcher labor saving, computer game-like environment, with *constrained sentence writing*, accrue the same advantages of the generative learning effect for L2 vocabulary learning, found in free sentence writing?

2) What is the relationship between performance on the inferencing computer games and performance on the vocabulary test?

2. Method

2.1. Pilot study

Consistent with iterative nature of design experiments (Gorard et al., 2004), the on-site coordinator of this research conducted a pilot study, and formative evaluation of the intervention instruments, using one reading passage and one computer game, with the idea of improving the materials before conducting a broader intervention.

Three students, two female and one male, from a Level B class (intermediate English proficiency) were recruited to participate in the pilot study at the small computer lab where classes are sometimes held.

In the pilot study, the onsite coordinator used formal protocol as a guidance to track the participants' behaviors when responding to usability tasks, and also to elicit their perceptions of the web-based text and computer game intervention. Specifically, the participants were invited to tell their first impressions and their perception of the purpose of the web-based text and computer game intervention, and then they were asked to conduct five tasks, and rate the usability of the program for each task. Their performances were observed and timed by the on-site coordinator. Afterwards, the three participants were invited to complete a short questionnaire of 17 Likert scale items, and 3 open-ended questions. The questionnaires were translated into Chinese to ensure the participants' full understanding of the items. Meanwhile, the on-site coordinator wrote observation reports and reflections based on five open-ended questions about the participants' reactions to several aspects of the IMapBook, including reading passages, new vocabulary definitions, feedback, and audio. After completing the pilot study, each participant was rewarded a small notebook.

Based on participant feedback, questionnaires and the onsite coordinator's observations of the participants interacting with the software, five major points emerged: (a) Generally, the participants liked learning English through the "computer-based games." The web-based text narrative and computer game intervention provided them with a new experience of learning vocabulary. (b) The three participants responded in very similar ways to the computer game, getting very similar inferences in the game. (c) The inference answers are not very satisfactory for them. As one pointed out, the inferences are rather rigid. (d) The text passage to read was difficult for them. (e) The audio was clear, but somewhat delayed due to the internet speed.

Based on these pilot study results, the research team and on-site coordinator made the following improvements in completing the materials: Easier text passages were used in the study. The research team made the computer games less rigid, by supplying more possible correct responses for the player. The on-site coordinator found a computer lab that had a faster internet connection, so that the audio would not be as delayed.

2.2. Context and participants

The study was conducted at a large comprehensive university in southwestern China. Fifty-seven EFL undergraduates from three level B College English classes participated in the study at a computer lab. Level B students have intermediate English proficiency, as opposed to high proficiency (level A), or low (level C), as described earlier. The participants' age range was 18–21.

College English is a required fundamental course (two years) for non-English major undergraduates in China. This course is composed of classroom-based instruction and computer-lab-based instruction with 75% of the time spent in classroom and 25% in the computer lab. The objective of the course is to develop students' ability to use the English language in a number of ways, including reading, using vocabulary in context, listening, speaking, writing, and intercultural communication. All the participants in the current study were in their second year of college English.

According to the National Chinese College English Curriculum, there are three levels of requirements, basic, intermediate, and higher requirements (described earlier). At Sichuan Normal University where the study took place, the students were enrolled in three different levels of English classes based on their scores on the National College Entrance English Exam and the University English Placement Test. Students with higher English proficiency were enrolled in the Level A class, students of lower proficiency were enrolled in the Level B class,

and those of lowest proficiency were enrolled in the Level C class. Level A classes implement higher requirements, Level B intermediate requirements, and Level C basic requirements.

2.3. Intervention and instrumentation

This study used a within-subjects design. Fifty-seven participants received both the experimental condition and the control condition, with the sequence counterbalanced. Each participant studied four text passages in the control condition and four in the experimental condition. The 57 participants were divided into two groups (29 in group 1 and 28 in group 2), The participants worked in two sessions of 2 h each.

In the control condition, students read booklets with (a) text passages containing some new vocabulary words, (b) a list of the new vocabulary words with their Chinese translations, (c) English definitions and the parts of speech. The students then answered three multiple-choice comprehension questions on the new vocabulary (also in the booklets). The control condition is a receptive learning condition (accent on reading, as opposed to writing), and incorporating relatively less generative learning effects (generating of associations between the new vocabulary word and long term memory) (Wittrock, 1974).

In the experimental condition, students read text passages online, with the new vocabulary words hyperlinked with glosses (popup definitions and the Chinese translation in Chinese characters). Following the reading, they played computer games involving making three inferences using the new vocabulary words. Fig. 1 shows a screen shot of an inference game.

The goal of the game was to make three valid inferences, based on the text passage, or story, preceding the game. When the player clicked on the buttons with words, a "click on word" interface (the lexicon; see the bottom panel in Fig. 1), a recording of a native English speaker saying the word played. The word was then placed in the panel currently labeled "*Your response will appear here…*" When the player felt their sentence was complete, they clicked on the "*Submit*" button and the program provided feedback on the validity of the inference in the context of the story, and also whether their sentence used one of the new vocabulary words (italicized in the lexicon). Players only earned points for inferences that contain at least one of the new vocabulary words and that were valid in the story context. If the attempt was valid, the player also received some elaborative feedback. So for example in the game shown in Fig. 1, clicking "*dogs,*" "*can,*" "*sniff,*" "*cancer*" and then on the "*submit*" button, earned one point and produced the feedback, "*True, dogs can be trained to distinguish cancer odors in patients. Terrific!*" When players entered sentences that were not in the set of valid inferences, they received feedback based on the pattern matching to the closest valid inference. So for example, if the player entered "*scientists,*" "*reward,*" "*noses,*" and clicked on the "submit" button, they received the feedback: "Did you mean *scientists reward* ____ ____ ____ ____ ?" This feedback is based on the nearest correct answer, i.e., "*scientists reward* dogs *for sniffing cancer.*" Based on the feedback, players made more attempts at generating inferences using the new vocabulary. They needed a total of three correct inferences, or three points, in each game to win.

The "computer games" used in the current study embodied to some extent the key elements of computer games, as defined by Malone (1981), Crawford (1984), and Gee (2003), i.e., (1) *rules* (or implied rules based on the game play structure): click on lexicons to form sentences that are inferences from the text, (2) *a start state*: (starting with no inferences generated), (3) *a goal for winning* (or set of win states): three inferences needed to win, (3) *immediate feedback on progress towards the goal*: (feedback on whether the inference is correct, and a unique qualitative response to most of the sentences possible), (4) *a game play space* (i.e., enough possible options in the *interaction* or play structure to give the player the perception of freedom of choice, playfulness or exploration): a large number of sentences possible to make with the lexicon, (5) *competition* (between two or more players, or between a single player and a computer opponent): limited in this case, and (6) *fantasy* (a storyline separate from the player's own life that allows them to experience another reality, without the real world risks of that reality "*psychosocial moratorium*"): to some extent the storyline from the text passage that precedes the game.

Note that the experimental condition, involving generating sentences (or inferences) in a highly constrained way, is designed to provide learners with a productive task (generating sentences, albeit in a highly constrained manner which requires less teacher and experimenter supervision). The experimental condition provides the learner with task involving generative learning tasks (generates associations
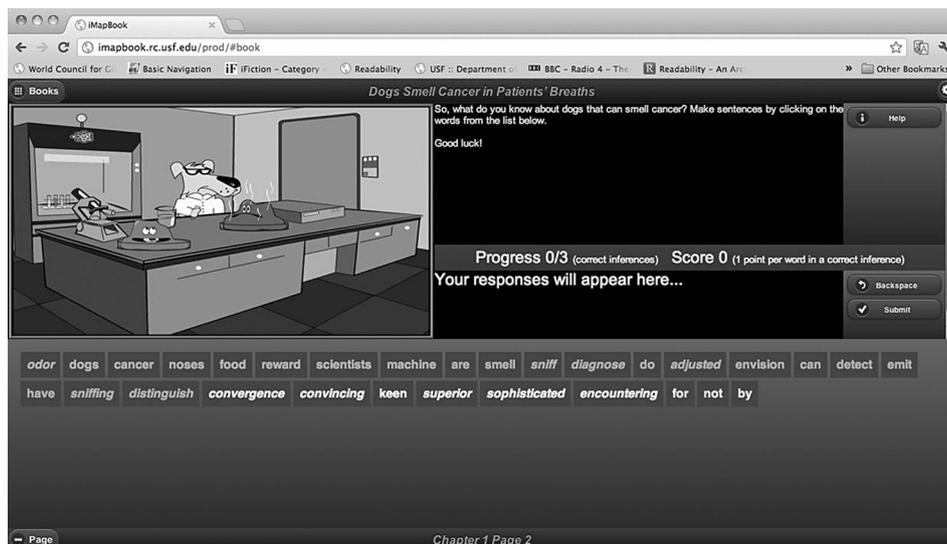

Fig. 1. Pretest and posttest, control is blue and experimental green.

between the new vocabulary word and the Long Term Memory (LTM)). Specifically the sentences that learner generates, since they are inferences about the story, connect with the passage just read, and with thus with anything in the passage with which the learner is familiar.

Participants full behavior, while reading the online text passages and playing the computer games, was recorded to a server-side database, where it was later accessed as part of the data. The text and game together are called "IMapBooks," and can be read and played as an interactive eBook on any device with a browser. IMapBook game refers to the computer authoring system in which graduate students in the southeastern university, with no knowledge of computer programming, created the computer games. The entire software suite, eReader for online text and associated computer games, authoring system for computer games, database, and reports, etc., is part of an infrastructure for embedding computer games into web-based eBooks, and conducting research on interactive reading, called *IMapBooks* (IMapBook.com).

2.4. Counterbalancing scheme

The study used a "*within-subjects*" design, meaning that all participants experienced all experimental conditions. The study used eight stories, two conditions, and two sessions of 2 h each for 57 participants. The study started with 60 participants (a generous sample sized for a within-subjects study), but lost three to attrition.

Each participant spent precisely the same amount of time (2 h) in each treatment condition. So there was no chance that results could be influenced by different times on task in the different conditions. Investigators divided each 2-h session in half and switched what the groups did in the second half of each session. Fig. 2 diagrammatically shows the arrangement. That is, in Session 1 in the first hour, Group 1 read two text passages online and played the associated computer games (experimental condition), and then in the second hour of the session Group 1 read two hardcopy booklets and answered the associated hardcopy multiple choice questions, etc. (control condition). The situation was reversed for Group 2. So, in Session 1 in the first hour, Group 2 studied two hardcopy booklets and in the second hour played two computer games. In Session 2, Group 1 started with hardcopy and then later switched to computer games. Also in Session 2, Group 2 started with computer games and later switched to hardcopy booklets. To remove the effect of any differences between text passages or sets of new vocabulary associated with a text passage, the order in which the eight text passages were read was also counterbalanced using a Latin square design (intuitively understandable through viewing Tables 1 and 2). Table 1 shows the scheme with time, or reading order, represented vertically, with the top earliest, and the bottom last, and each column representing approximately one eighth of the participants. The first participant, in the leftmost column of Table 1, read in this order: passage 1, passage 2, …, up to passage 8. The second participant read passage 8, passage 1, …, passage 7. The eighth participant, in rightmost column, read passage 2, passage 3, …, passage 8, passage 1. With the ninth participant, the cycle starts over. The ninth participant, in leftmost column, read in this order: passage 1, passage 2, …, up to passage 8.

2.5. Procedures

Fig. 3 shows the time sequence for the study. The students took an orientation to the study before participating in the two learning sessions. The students who were willing to participate signed the consent forms. Before the intervention, students took a questionnaire on basic information about English learning and computer use. The participants were also pretested on the new vocabulary words with the Vocabulary Knowledge Scale (VKS) (Paribakht & Wesche, 1993).

Next, participants were randomly assigned to two groups for counterbalancing. Both groups learned vocabulary in both of the two conditions, as was described in Section 2.2. During the intervention, all the students' game playing behaviors were automatically recorded in a log file on a server-side database.

On the day the students completed the second learning session, they ended with questionnaires about their perceptions of the computer games (including five-point Likert-scale questionnaire items and short-answer questions). Six students also volunteered for individual

Session 1

	Cohort 1	Cohort 2
Time 1	X	O
Time 2	O	X

Session 2

	Group 1	Group 2
Time 1	O	X
Time 2	X	O

Fig. 2. Cohort-session counterbalancing scheme: X is two online text passages plus computer games; O is two hardcopy text passages plus multiple-choice questions.

Table 1
Counterbalancing scheme for the order participants received stories (story 1 through story 8).

1	8	7	6	5	4	3	2
2	1	8	7	6	5	4	3
3	2	1	8	7	6	5	4
4	3	2	1	8	7	6	5
5	4	3	2	1	8	7	6
6	5	4	3	2	1	8	7
7	6	5	4	3	2	1	8
8	7	6	5	4	3	2	1

semi-structured interviews. However, the questionnaires and interviews are not included in the current paper (which focuses on the quantitative data) but in a follow-up qualitative paper.

In order to assess new vocabulary learning and retention, 3 days after the experimental sessions, participants again took the VKS posttest (with the same test items and format as the pretest). On the same day, each participant who completed all the tasks in the study was rewarded with a small gift. Analyses focused on two main data sources: preliminary and follow-up vocabulary tests and logged game performance data.

Each participant encountered all of the 40 words in the pretest, posttest and during the intervention. However, depending on where they fell in the counterbalancing scheme, they encountered 20 of the new vocabulary words in the experimental condition and 20 in the control condition.

2.6. Data analysis

To examine the effectiveness of the two learning conditions, VKS pre- and posttests were analyzed. Each VKS test was independently checked by two graduate students (or two investigators). Test items with discrepancies were later resolved by the full group (all graduate students and investigators).

Following is an example completed VKS item from the study:
"Reconstruct:

1. I have never seen this word.
2. I have seen this word before, but I don't know what it means.
3. I have seen this word before, and I think it means _____ . (synonym or translation)
4. I know this word. It means ____ *form again*/重建 ____ (synonym or translation)
5. I can use this word in a sentence: *The destroyed highway is being reconstructed.*"

Since participants could mark more than one answer for a question, the researchers decided that the highest verifiably correct answer would be used as the data point. Note that answers "1" and "2" are not verifiable, while "3," "4," and "5" are verifiable. Therefore, interpretation was necessary for answers in the "3" to "5" range, but not if "1" or "2" was the highest answer. The first step was to check answers in the "3" to "5" range. In other words, graduate students who were native English speakers judged whether a participant's synonym or translation (level 3 or 4 answers) was indeed valid. Answers that included Chinese translations (Chinese characters) were first translated to English by a native Chinese speaker (who also judged the validity of the answer) and then judged in the English language version by native speakers. English speakers judged whether the student correctly used the supplied vocabulary word in context (level 5). Next the highest level answer was marked with a stamp. Finally, the data, the stamped values, were entered into an Excel file, and later uploaded to the statistical program SPSS for analysis.

To explore the participants' performance in the inferencing computer games, the investigators analyzed the log files in the database, which contained every response that participants made in the games. The number of correct inferences per game for each participant was downloaded to an Excel file, and later uploaded to SPSS.

2.6.1. Sorting out pre- and post-test VKS scores, according to the counterbalancing scheme

Given the within-subject design, and the complexity of the counterbalancing scheme (experimental versus control, and order of the narratives presented), it was an exacting process to figure out what constituted the experimental conditions for the pretest and the posttest for each participant. The Pre- and Post-test were identical form of VKS items covering 40 vocabulary items. However, because of the counterbalancing, there were eight variations of which new vocabulary items participants encountered during treatment in control versus experimental conditions. However, ultimately each participant encountered 20 of the total 40 new vocabulary words in the experimental

Table 2
Descriptive statistics of pre- and post-test.

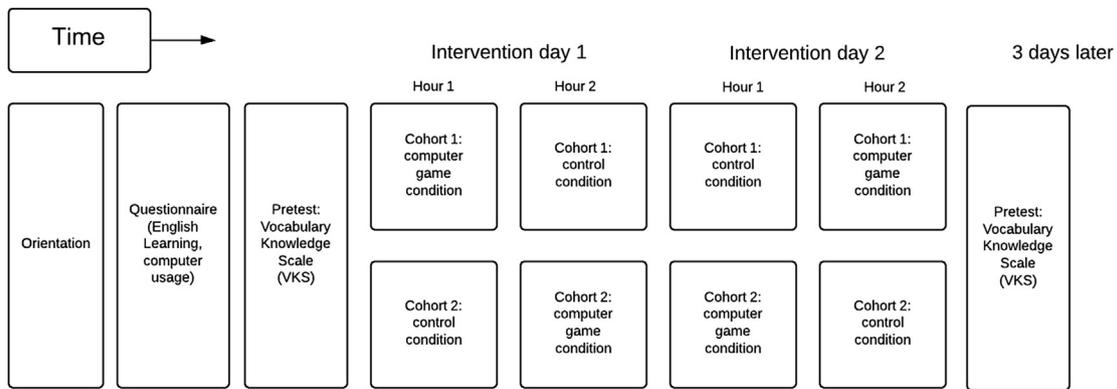| Level of difficulty	Standard		
	Mean error rate	Deviation	Sample size
Pretest Control	1.81	.399	54
Pretest IMapBook	1.84	.423	54
Posttest Control	2.66	.645	54
Posttest IMapBook	3.02	.656	54

Fig. 3. The time sequence for the study.

condition and 20 in the control condition. Thus for each participant, for both pre- and post-test, during analysis of data, the posttest data was divided into 20 words encountered in the experimental condition and 20 words in the control condition. For each participant, 20 new vocabulary items corresponded to pre-test and post-test *control condition*, and the remaining 20 new vocabulary items corresponded to pre-test and post-test *experimental condition*.

3. Results

3.1. Research question 1

What is the effectiveness of the inferencing computer games compared with the hardcopy booklets for vocabulary retention? (Can an automated, labor saving, computer game-like environment, with constrained sentence writing, accrue the same advantages of the generative learning effect for L2 vocabulary learning, found in free sentence writing?)

Fig. 4 graphically summarizes the results from the VKS vocabulary tests, while Table 3 summarize them in terms of means, standard deviations, significance sizes, and effect sizes.

In the pretest, there was no significant difference between control ($M = 1.81$, $SD = .399$) and IMapBook ($M = 1.84$, $SD = .423$), as indicated by t-test, $t(1,53) = .617$. The differences between the posttests, IMapBook ($M = 3.026$, $SD = .656$) and control ($M = 2.67$, $SD = .645$) were significant, $t(1, 53) = 4.09$, $p < .0005$, $d = .56$, with a medium effect size. While these t-tests are easily understood, a more correct analysis examines the pre- and posttests dynamically over time (see below).

As noted in the earlier discussion, each participant experienced the pretest (with all the words), both conditions of learning (control and IMapBook/computer game each with half the words), followed by the posttest. For each participant, the words encountered in the IMapBook condition were different words than those encountered in the hardcopy booklets condition; thus, the data of pretest and posttest for each participant were separated into pretest control and pretest IMapBook and posttest control and posttest IMapBook.

Means and standard deviations for the VKS pre- and posttests of the new vocabulary words are shown in Table 2. The means are averages of the answers in the scale of one to five used in VKS questions, with "1" indicating the least knowledge of a vocabulary word and "5" the greatest. As Table 2 shows, under the control condition, the mean of vocabulary knowledge in the posttest ($M = 2.66$, $SD = .645$) is greater
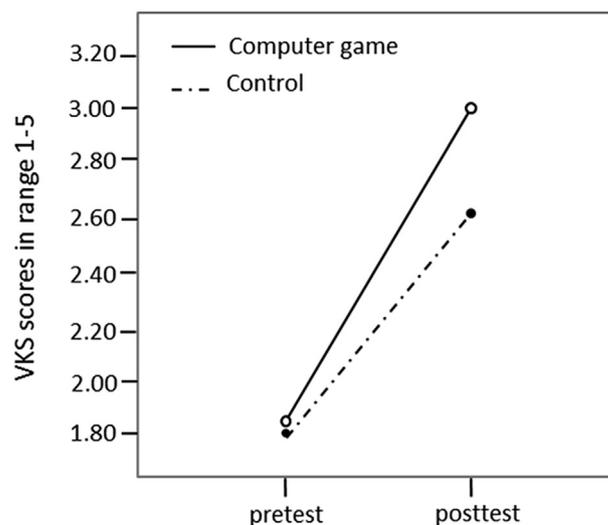

Fig. 4. Mean scores for pretest and posttest VKS, control is dotted and experimental solid.

Table 3
Statistical significances and effect sizes for VKS vocabulary between condition tests.

Statistical test	Significance level	Effect size
T-test pretests	$t(1,53) = .617, p < .617$	None
T-test posttests	$t(1, 53) = 4.09, p < .0005$	$d = .56$, medium
ANOVA main effect between conditions	$F(1, 53) = 9.37, p < .003$	Partial Eta Squared = .15, medium
ANOVA interaction between time and condition	$F(1, 53) = 19.94, p < .0005$	Partial Eta Squared = .27, large
ANOVA main effect of time	$F(1, 53) = 360.90, p < .0005$	Partial Eta Squared = .872, large

than the mean in the pretest ($M = 1.81$, $SD = .399$). Also, the mean of vocabulary knowledge in the posttest for the experimental condition ($M = 3.02$, $SD = .656$) is larger than the mean in the pretest ($M = 1.84$, $SD = .423$). That is, both the hardcopy booklet and the inferencing computer games procedures led to higher scores in the vocabulary retention posttest, compared with pretest performance. There was a greater difference between pre- and posttests in the experimental condition (IMapBook with computer games) than in the control condition (hardcopy with multiple choice questions).

In order to test whether these between-conditions differences in pre- versus posttests were statistically significant, the investigators ran a within subjects analysis of variance. The analysis yielded a significant main effect of time, that is, pre- to posttest across condition, $F(1, 53) = 360.90$, $p < .0005$, partial eta squared is .872. The analysis also yielded a significant main effect of condition (control versus experimental), combining pre- and post-test scores, $F(1, 53) = 9.37$, $p < .003$. The partial eta squared of .15 suggests a medium effect size (Cohen, 1988). Given that there was no significant difference between conditions on the pretest, and there was a significant difference on the posttest (see above), the condition effect is attributable to a greater difference pre- to post- in the experimental condition, than in the control condition. Most telling is the significant interaction between time and condition, reflecting differences in pre to posttest change between the two conditions, $F(1, 53) = 19.94$, $p < .0005$. A partial *eta* squared of .27 reflects a large effect size (Cohen, 1988). This result is due to a greater increase in pre- to post scores in the computer game condition than in the control condition. Table 3 shows the various significance levels and effect sizes.

3.2. Research question 2: what is the relationship between performance on the inferencing computer games and performance on the vocabulary test?

The hypothesis that more correct inferences during the gaming would be associated with higher scores in the VKS vocabulary posttest was supported by correlational analysis. Analysis yielded a strong correlation between the number of correct inferences and the pre- to posttest gain on the VKS for the IMapBook, $r(33) = .515$, $p < .01$. A subset of participants ($n = 34$) was used for this analysis because because some participants' names handwritten on the hardcopy VKS vocabulary tests could not be connected with their typed in names in the computer game database.

Since every response that participants made in the inference games was recorded in a database, analysis of the log files also provided an overview of the participants' inferencing-game behaviors. Generally, in the inference games, participants submitted an average of 16.2 attempts at inferences per game with 11.3% of them correct, which is rather low. With responses classified as valid (1.0) or invalid (.0), the mean was .113 and the standard deviation was .316. On average, the participants got 1.83 correct inferences per game. They needed to generate three correct inferences to win each game. That is, those who completed a total of twelve correct inferences won all of the four inferencing computer games.

4. Discussion

4.1. Interpretation of results

The results indicate that inference-based computer games result in better learning of new vocabulary than standard rote-memorization vocabulary practices that use hardcopy lists of new vocabulary words and multiple-choice questions. The better vocabulary posttest results for the gaming/inferencing condition suggest that gaming has the potential for more and better vocabulary learning. Further, the significant correlation between the number of correct inferences in the game and the score in the vocabulary posttest in the gaming condition is consistent with the proposal that achievement in the game can predict improved vocabulary learning.

The current study is a design experiment (Cobb et al., 2003; Reinking & Bradley, 2008) that investigates an intervention for a specific situation (Chinese undergraduates studying English vocabulary). As such, it demonstrates that a computer game approach is an attractive alternative to hardcopy text, list of words and multiple-choice questions, for Chinese undergraduate students to study English vocabulary. But such questions can be investigated at multiple levels of granularity. Is it the scoring, and motivational qualities, of the computer game that provide the benefit, or the immediate feedback, or is it reading online versus hardcopy? The current investigators take the position that computer games have identities as an object. Players have a "computer game" schema. Before you isolate, and investigate one factor at a time, all the individual factors that make up "computer game," it makes sense to investigate the bundle of factors as a computer game. By isolating factors of computer games, you will likely lose some of the emergent qualities of computer games, as Malone (1981) did when he did experiments isolating which features of computer games are most important for motivation. Therefore, current investigators provide a comparison with a "computer game" condition, acknowledging any computer game is made up of many factors, plus emergent qualities of all these factors in combination. We leave the isolation of factors to other follow up studies.

Our results support the proposal that (compared with standard hardcopy booklets) inference-based computer games lead to deeper processing of vocabulary, resulting in better recall. This is consistent with predictions made within a levels of processing framework (Craik & Lockhart, 1972). That is, the elaborative process required for making inferences results in deeper, more effective encoding, compared to reading lists of words and doing multiple-choice questions pro forma, as in the hardcopy condition of this study.

The current results suggest that automated and constrained sentence generation activities using limited lexicons with a "click on word" interface, in a computer game-like setting, can lead to generative learning advantages (generation of associations between new words and LTM) for L2 new vocabulary learning, similar to those found by Meyers (2010), but without the logistical problems of having a human verify whether the sentences the students generated conform to generative learning specifications.

The current study also demonstrates that use of a computer game intervention, customized to a formal learning situation can lead to significant vocabulary learning gains in a short time (two sessions of 2 h). Cobb & Horst (2011) had participants, 11–12 years old, play with the off the shelf computer game, Word Coach, for two months to produce significant gains in vocabulary. Because of multiple differences in method (participant age, etc., invention type), we do not make a direct comparison. However, it is worth noting that computer games, designed for the specific situation, can make learning gains in a short period of time.

The inference computer games also provide an attractive alternative for L2 learners to study vocabulary. The current study, as a design experiment, demonstrates a more effective way for the specific target audience, undergraduate Chinese English second language learners, to study new vocabulary.

Additional factors, along with inferencing, that may have contributed to student motivation and to learning in the gaming condition are the feedback, pictures, and voice of a native speaker speaking the lexicon words. Multiple factors and modalities may create more memory connections and thus result in better learning of new vocabulary (Groot, 2000; Ma & Kelly, 2006). Multimedia applications, in general, and computer games, in particular, are well suited to bundling these multiple modalities (Chun & Plass, 1996).

Immediately prior to the study, the investigators thought that the computer games would be too difficult for the Chinese EFL students. Because the existing infrastructure made it almost impossible for the designers to include every possible valid inference, there were many valid inferences that the game did not accept as valid. The investigators did not retroactively classify those inferences as valid, for data analysis, because they were conducting a design experiment to investigate the practicality of a game approach, including difficulty of administration. Additionally, if inferences were retroactively classified as valid, after participants had received feedback during the experiment that they were invalid, that would create an inconsistency.

The small number of valid inferences in the system made the games, in the opinion of the designers, excessively hard. However, the investigator who coordinated the study onsite in China reported that participants enthusiastically tackled the games and were happy to have this innovative means of studying vocabulary. Difficult games did not seem to discourage the Chinese EFL college students in vocabulary learning. This observation echoed previous research findings that with the norms of computer game play, people accept and embrace a higher level of challenge than they would in the classroom (Gee, 2003, 2007, 2008).

4.2. Limitations

The same VKS vocabulary test was used for both the pretest and the posttest. In general, people improve their scores by merely retaking the same test (the test–retest practice effect), even without any learning intervention (Collie, Maruff, Darby, & McStephen, 2003). In the current study, both conditions resulted in significant improvements from pre- to post-test. It is difficult to sort out how much of this improvement results from retaking the same test and how much from the interventions. It is, however, clear that the eBook with computer game condition resulted in more vocabulary learning than the traditional hardcopy lists of words and multiple-choice questions.

4.3. Implications for pedagogy, instructional design and research design

The significant correlation between game scores and the vocabulary posttest suggests the possible development of game-based stealth assessments of vocabulary learning. Such game-based assessments could be developed and calibrated to have concurrent validity (Beasley, Jason, & Miller, 2012; Cronbach & Meehl, 1955; Jeremy, 2004).

The current study used an integrated interactive eBook system (IMapBooks.com) with authoring system designed to embed computer games in eBooks, a database to automatically record students' game-play behavior and a report system to supply the researchers with game-play behavior summaries. Graduate students with no technical knowledge developed the materials (text passages followed by computer games). Authoring systems to create interactive eBooks, are likely to become increasingly available as the education sector becomes aware of interactive eBook's potential. This suggests the possibility that educators or school librarians might also create custom interactive eBooks for education, adapted to learning standards. Because these interactive eBooks systems can record game-playing into a database and later supply summary reports on player behaviors, they have the potential for research on literacy and reading.

5. Conclusion

Chinese college students in EFL courses learned more new vocabulary using web-based eBooks with inference-based computer games than they did with more traditional methods (hardcopy readings, word lists, and multiple-choice questions). Further, their game scores were significantly correlated with the amount of vocabulary learned, suggesting that motivated game play and game achievement were causal factors in the learning. Gaming *as part of* studying motivates students to practice and learn new vocabulary and often challenges educators to create innovative ways of teaching and learning second and foreign language, particularly in the EFL context. It also challenges educators to connect gaming to the main curriculum for EFL learners in the dynamic global world. If we ask whether college students should *play games or study*, in this case the answer seems to be that college students should play games *to* study.

References

Abraham, L. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: a meta-analysis. *Computer Assisted Language Learning, 21*, 199–226.

AbuSeileek, A. F. M. (2008). Hypermedia annotation presentation: learners' preferences and effect on EFL reading comprehension and vocabulary acquisition. *CALICO Journal, 25*, 260–275.

Anderson, J. R., & Reder, L. M. (1979). An elaborative processing explanation of depth of processing. In L. S. Cermak, & F. I. M. Craik (Eds.), *Levels of processing in human memory*. Hillsdale, N.J: Erlbaum.

Başoğlu, E. B., & Akdemir, . (2010). A comparison of undergraduate students' English vocabulary learning: using mobile phones and flash cards. *The Turkish Online Journal of Educational Technology, 9*, 1–7.

Beasley, C. R., Jason, L. A., & Miller, S. A. (2012). The general environment fit scale: a factor analysis and test of convergent construct validity. *American Journal of Community Psychology, 50*(1–2), 64–76. http://dx.doi.org/10.1007/s10464-011-9480-8.

Birdsong, D., & Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of Memory and Language, 44*, 235–249.

Bradshaw, G. L., & Anderson, J. R. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior, 21*, 165–174.

Bytheway, J. (2011). *Vocabulary learning strategies in massively multiplayer online role-playing games*. Victoria University of Wellington. Unpublished masters thesis.

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

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal, 80*(2), 183–198. http://dx.doi.org/10.1111/j.1540-4781.1996.tb01159.x.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Collie, A., Maruff, P., Darby, D. G., & McStephen, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. *Journal of the International Neuropsychological Society, 9*(3), 419–428. http://dx.doi.org/10.1017/S135561770393007L.

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

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671–684. http://dx.doi.org/10.1016/S0022-5371(72)80001-X.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268–294.

Crawford, C. (1984). *The art of computer game design*. Berkeley, CA: Osborne/McGraw-Hill.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. http://dx.doi.org/10.1037/h0040957.

deHaan, J., Reed, W. M., & Kuwada, K. (2010). The effect of interactivity with a music video game on second language vocabulary recall. *Language Learning & Technology, 14*(2), 74–94.

Dickey, M. D. (2011). Murder on Grimm Isle: the impact of game narrative design in an educational game-based learning environment. *British Journal of Educational Technology, 42*(3), 456–469.

Ellis, N. (1995). The psychology of foreign language vocabulary acquisition: Implication of CALL. *Computer Assisted Language Learning, 8*(2), 103–128. http://dx.doi.org/10.1080/0958822940080202.

Flege, J. E., Jeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second language acquisition. *Journal of Memory and Language, 41*, 78–104.

Gan, Z., Humphreys, G., & Hamp-Lyons, L. (2004). Understanding successful and unsuccessful EFL students in Chinese Universities. *The Modern Language Journal, 88*(2), 229–244.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Gee, J. P. (2007). *Good video games + good learning*. New York: Peter Lang Publishing Inc.

Gee, J. P. (2008). Learning and games. In K. Salen (Ed.), *20. The ecology of games: Connecting youth, games, and learning, the John D. and Catherine T. MacArthur foundation series on digital media and learning* (pp. 21–40). Cambridge, MA: The MIT Press.

Gettys, S., Imhof, L. A., & Kautz, J. O. (2001). Computer-assisted reading: the effect of glossing format on comprehension and vocabulary retention. *Foreign Language Annals, 34*, 91–106.

Gorard, S., Roberts, K., & Taylor, C. (2004). What kind of creature is a design experiment? *British Educational Research Journal, 30*(4), 577–590.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371–395.

Groot, P. J. M. (2000). Computer assisted second language vocabulary acquisition. *Language Learning & Technology, 4*(1), 60–81.

Gu, Y., & Johnson, R. (1996). Vocabulary learning strategies and language learning outcome. *Language Learning, 46*(4), 643–679.

Haynes, M. (1990). Examining the impact of L1 literacy on reading success in a second writing system. In H. Burmeister, & P. L. Rounds (Eds.), *Variability in second language acquisition: Proceedings of the tenth meeting of the second language research forum*. Eugene, OR: Department of Linguistics, University of Oregon.

Hwang, H., Shadiev, R., & Huang, S.-M. (2012). A study of a multimedia annotation system and its effect on the EFL writing and speaking performance of junior high school students. *ReCALL, 23*, 160–180.

Jeremy, E. C. (2004). The index of learning styles: an investigation of its reliability and concurrent validity with the preference test. *Individual Differences Research, 2*(3), 169–174.

Jia, G., Aaronson, D., & Wu, Y. (2002). Long-term language attainment of bilingual immigrants: predictive variables and language group differences. *Applied Psycholinguistics, 23*, 599–621. DOI: 10.1017.S0142716402004058.

Kayaoglu, M. N., Dag Akbas, R., & Ozturk, Z. (2011). A small scale experimental study: using animations to learn vocabulary. *The Turkish Online Journal of Educational Technology, 10*, 24–30.

Kispal, A. (2008). *Effective teaching of inference skills for reading. Literature review*. Research report DCSF-RR031. Berkshire, UK: National Foundation for Educational Research.

Lyman-Hager, M. A., Davis, J. N., Burnett, J., & Chennault, R. (1993). Une vie de boy: interactive reading in French. In F. L. Borchardt, & E. M. T. Johnson (Eds.), *Proceedings of the CALICO 1993 annual symposium on "assessment"* (pp. 93–97). Durham, NC: Duke University.

Ma, Q., & Kelly, P. (2006). Computer assisted vocabulary learning: design and evaluation. *Computer Assisted Language Learning, 19*(1), 15–45. http://dx.doi.org/10.1080/09588220600803998.

Malone, T. W. (1981). Toward a theory of intrinsically instruction motivating. *Cognitive Science, 4*, 333–369. doi: 10.1.1.103.6313.

Mandler, G., & Dorfman, J. (1994). Implicit and explicit forgetting: when is gist remembered? *The Quarterly Journal of Experimental Psychology*, 651–672.

Martinez-Lage, A. (1997). Hypermedia technology for teaching reading. In M. Bush, & T. Terry (Eds.), *Technology enhanced language learning* (pp. 121–163). Lincolnwood, IL: National Textbook.

Meyers, P. C. (2010). *Incidental foreign language vocabulary learning from generative tasks*. USA: Temple University. Unpublished doctoral thesis.

Miller, M., & Hegelheimer, V. (2006). The SIMS meet ESL: incorporating authentic computer simulation games into the language classroom. *International Journal of Interactive Technology and Smart Education, 3*(4), 311–328.

Mohsen, M. A., & Balakumar. (2011). A review of multimedia glosses and their effects on L2 vocabulary acquisition in CALL literature. *ReCALL, 23*, 135–159.

Mondria, J. A. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings: an experimental comparison of the 'meaning-inferred method' and the 'meaning-given method'. *Studies in Second Language Acquisition, 25*, 473–499.

Mondria, J.-A., & Wiersma, B. (2004). Receptive, productive, and receptive + productive L2 vocabulary learning: what difference does it make? In P. Bogaards, & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 79–100) Amsterdam/Philadelphia: John Benjamins.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. New York: Cambridge University Press.

Oberg, A. (2011). Comparison of the effectiveness of a CALL-based approach and a card-based approach to vocabulary acquisition and retention. *CALICO Journal, 29*, 118–144.

Paribakht, T. S., & Wesche, M. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal, 11*(1), 9–29.

Ranalli, J. (2008). Learning English with the Sims: exploring authentic computer simulation games for second language learning. *Computer Assisted Language Learning, 21*(5), 441–455. http://dx.doi.org/10.1080/09588220802447859.

Rankin, Y. A., Gold, R., & Gooch, B. (2006). 3D role-playing games as language learning tools. In , *Vol. 25. Paper presented at the EuroGraphics 2006*. Vienna: Austria. September 4–8, 2006 http://www.thegooch.org/.

Rankin, Y. A., Morrison, D., McNeal, M., Gooch, B., & Shute, M. W. (2009). Time will tell: In-game social interactions that facilitate second language acquisition. In R. Young (Ed.), *Proceedings of the 4th international conference on foundations of digital games* (pp. 161–168). New York: ACM.

Reinking, D., & Bradley, B. A. (2008). *On formative and design experiments*. New York: Teachers College Press.

Roussel, S. (2011). *A computer assisted method to track listening strategies in second language learning*.

Segler, T., Pain, H., & Sorace, A. (2002). Second language vocabulary acquisition and learning strategies in ICALL environments. *Computer Assisted Language Learning, 15*(4), 409–422.

Sewell, E. H. (2008). Language policy and globalization. In E. Peterson (Ed.), *Communication and public policy proceedings of the 2008 international colloquium of communication* (pp. 74–80).

Shaunessy, M., & Dinnell, D. (1999). Levels of elaboration, interference and memory for vocabulary definitions. *North American Journal of Psychology, 1,*(2), 293–306.

Smith, R. (2005). Global English: gift or curse? *English Today, 21*(2), 56. http://dx.doi.org/10.1017/S0266078405002075.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(5), 295–312. http://dx.doi.org/10.1016/0959-4752(94)90003-5.

Thorne, S. T., Fischer, I., & Lu, X. (2012). The semiotic ecology and linguistic complexity of an online game world. *ReCALL, 24*(3), 279–301.

Trabasso, T., & Magliano, J. P. (1996). Conscious understanding during comprehension. *Discourse Processes, 21*, 255–287.

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist, 11*(1), 87–95.

Wittrock, M. C. (1990). Generative processes of comprehension. *Educational Psychologist, 24*(4), 345–376.

Wu, X., Lowyck, J., Sercu, L., & Elen, J. (2013). Task complexity, student perceptions of vocabulary learning in EFL, and task performance. *British Journal of Educational Psychology, 83*, 160–181.

Young, M., Slota, S., Cutter, A., Jalette, G., Mullin, G., Lai, B., et al. (2012). Our princess is in another castle: a review of trends in serious gaming for education. *Review of Educational Research, 82*, 61–89. http://dx.doi.org/10.3102/0034654312436980.

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: a meta-analysis. *Computer Assisted Language Learning, 24*, 39–58.

Zheng, D., Young, M., Wagner, M., & Brewer, R. (2009). Negotiation for action: English language learning in game-based virtual worlds. *Modern Language Journal, 93*, 489–511. http://dx.doi.org/10.1111/j.1540-4781.2009.00927.x.