Digital game-based learning’s effectiveness on EFL learners’ receptive and productive vocabulary knowledge

Wen Jia, Nanjing Agricultural University; Xi’an Jiaotong-Liverpool University
Liping Zhang, Army Engineering University of PLA
Austin Pack, Brigham Young University-Hawaii
Yi Guan, Global Institute of Software Technology
Bin Zou, Xi’an Jiaotong-Liverpool University

Abstract

Although digital game-based vocabulary learning (DGBVL) has received increasing attention in the past two decades, the impacts of DGBVL on the depth of word knowledge are still not well understood, especially in regard to productive vocabulary learning and DGBVL’s long-term efficacy. This study leverages a quasi-experimental research design to investigate DGBVL’s long-term effects on receptive vocabulary (RV) and productive vocabulary (PV). Forty-eight Chinese English-as-a-foreign-language (EFL) university students, assigned to the experimental and control groups, were instructed by a DGBVL approach and PowerPoint (PPT) lecturing, respectively, over the course of 18 weeks. Specifically, a mixed 2×2 repeated measures experimental design was conducted by adopting instruction type (DGBVL and PPT lecturing) and testing time (pretest and posttest) as the independent variables, with RV and PV proficiency as the respective dependent variables. The results suggest that instruction type and teaching time have significant effects on participants’ RV and PV learning achievements. However, teaching time’s effect size outweighs instruction type. The findings are highly encouraging for the use of DGBVL in the EFL classroom, as it may serve as an effective and long-lasting pedagogical tool within this context.

Keywords: Digital Game-Based Learning, Receptive Vocabulary, Productive Vocabulary, English as a Foreign Language

Language(s) Learned in This Study: English


Introduction

While “vocabulary knowledge is most likely to develop if there is a balance of incidental and deliberate appropriate opportunities for learning” (Nation, 2020, p. 15), many English-as-a-foreign-language (EFL) learners, such as those in China, lack opportunities for exposure to the target language (i.e., English), thereby making incidental vocabulary learning challenging in formal and informal settings. In such cases, intentional vocabulary learning in EFL learners’ schooling serves as a valuable and complementary approach (Webb, 2020) by providing more input and output practices through direct instruction. The teacher-centered lecturing and focus on rote memorization of vocabulary have long dominated China’s EFL context (Yang & Dai, 2011), resulting in tedious and inefficient lexical learning. Encouraging, however, are emerging cutting-edge technologies that afford new avenues of vocabulary learning, including the use of digital games.
While the idea that games, playing, and learning all interrelate extends back into ancient history (Pack & Newbould, 2018), the educational value of digital games has received growing attention since Prensky's (2001) comprehensive description of digital game-based learning (DGBL). Digital games have been acknowledged as potential pedagogical tools with applications in a variety of fields. It was not until advances in technology and digital games became more accessible in the mid-2000s that interests in digital game-based language learning (DGBLL) were kindled in the field of second and foreign language teaching (Reinhardt, 2019). This form of language pedagogy has drawn attention because it is able to afford opportunities for edutainment (i.e., a blend of education and entertainment), and may assist in engaging and motivating learners, as well as in facilitating increased interaction and exposure to the target language (Reinders, 2012). Compared with studies investigating language reading (e.g., Dourda et al., 2014), speaking and listening (e.g., Hwang et al., 2016), writing (e.g., Allen et al., 2014), grammar (e.g., Reitz et al., 2016), pronunciation (e.g., Cerezo et al., 2019) and other language skills, digital game-based vocabulary learning (DGBVL) has drawn the most attention to date in the DGBLL field.

Studies exploring DGBVL do so from three major angles: its affordances (e.g., Rama et al., 2012); the cognitive and psychological consequences, such as motivation, engagement, and anxiety (e.g., Yang et al., 2020); and most commonly, its influence on vocabulary learning (e.g., Alfadil, 2020; Rasti-Behbahani & Shahbazi, 2020). Studies on the latter have mostly been carried out in formal settings (i.e., L2 classrooms) and report the positive effects of DGBVL. Several studies have also observed the effects of DGBVL in extramural learning environments, with frequent gamers outperforming moderate and non-gamers (e.g., Sylvén & Sundqvist, 2012).

Though some research has been carried out on DGBVL, much uncertainty still exists about the relation between DGBL and its efficacy for vocabulary knowledge learning. First, most of the extant literature focuses either on vocabulary in very general terms (e.g., Ranalli, 2008; Wu, 2018) or on receptive vocabulary in particular (e.g., Park et al., 2019; Tseng et al., 2020); only a handful of studies make an explicit distinction among vocabulary knowledge types and draw specifically on productive vocabulary (deHaan et al., 2010; Franciosi, 2017) or both receptive vocabulary (RV) and productive vocabulary (PV) (Rasti-Behbahani & Shahbazi, 2020; Sundqvist & Wikström, 2015). Second, DGBL treatment time is another considerable factor. As Zou et al. (2019) note, although a consensus of positive short-term (i.e., tens of minutes to several days) effectiveness of DGBVL has more or less been achieved, disagreement remains as to the long-term effects of DGBVL. Third, previous research on long-term incidental vocabulary acquisition in informal settings (e.g., Sundqvist, 2019; Sundqvist & Wikström, 2015) does not elucidate the subtle changes of RV and PV that occur in intentional vocabulary learning over time. Therefore, this 18-week longitudinal study aimed to investigate EFL learners’ intentional RV and PV learning efficacy from DGBVL in the formal setting as well as explore DGBVL’s long-term effects.

Related Conceptions and Literature Review

Receptive Vocabulary and Productive Vocabulary

Nation developed the word knowledge process model describing what is involved in knowing a word in 2001, and now the model is most widely known in vocabulary studies (Nation, 2020; Webb, 2013). In this model, Nation divides the vocabulary learning process, be it receptive or productive vocabulary, into three aspects: form, meaning and use. Receptive vocabulary refers to vocabulary items whose form, meaning, and use can be learned while listening or reading; productive vocabulary involves the ability to use a word in the right form, meaning, and usage in speech or writing (Nation, 2001). But in terms of language use, “the most important aspects of vocabulary knowledge for an EFL learner are knowledge of word form and the form-meaning connection” (Nation, 2020, p. 18). Considering the EFL context and feasibility of this study, the current study’s working definitions and measurement dimensions of RV and PV follow Nation’s (2001) model, but solely focus on the written form of RV and PV at the form-meaning link.
DGBVL for Receptive Vocabulary Learning

Studies have explored RV learning yielded from digital games from a variety of perspectives. The first cohort of studies compares game and non-game approaches. These frequently report various games’ influences on vocabulary learning by comparing the DGBL method with other traditional vocabulary learning approaches. Example games include video games (e.g., Calvo-Ferrer, 2017; Chen & Yang, 2013), serious games (e.g., Chen & Hsu, 2020), gamified vocabulary applications (e.g., Dindar et al., 2021; Wu, 2018), virtual games (e.g., Cerezo et al., 2019), amongst others. Some earlier studies explored video games’ impacts on vocabulary learning both in formal settings (Calvo-Ferrer, 2017) and informal settings (Chen & Yang, 2013). Findings suggest that video games tend to be more helpful in the short run than in the long run. Some of these conclusions have been questioned, given they are drawn from students’ subjective self-reporting data (Chen & Yang, 2013). More recently, studies have turned to virtual games that afford more advanced technological support. For example, the use of VR (Alfadil, 2020) and 3D games (Tseng et al., 2020) has been compared with traditional teaching approaches. Both cited experimental studies witnessed significantly higher vocabulary immediate posttest scores than pretest scores in experimental groups that leveraged virtual environments, suggesting that virtual games could facilitate vocabulary learning by providing an interactive and contextualized situation for vocabulary presentation.

The second category of studies examines vocabulary learning outcomes resulting from different game designs or gamified functions within the same game. Several have explored the efficacy of games’ reward systems, such as providing badges and/or displaying leader boards to users based on particular achievements. For example, Park et al. (2019) explored the effects of a performance-contingent reward mechanism on participants’ learning, motivation, engagement, and system perception. They found that the reward mechanism produced a statistically significant increase in the level of participants’ receptive vocabulary learning, motivation, and engagement compared with completion-contingent rewards in the same kind of game. Features such as competition and cooperation have also drawn researchers’ attention. However, unlike more positive results demonstrated in the game and non-game approach comparisons, some studies found no significant difference in students’ vocabulary learning between cooperative and competitive conditions (Dindar et al., 2021; Peng et al., 2016).

The third group of studies makes use of a cohort study design rather than a controlled trial in order to observe the vocabulary gains or retention that result from DGBVL intervention, and to shed light on the nuances of the process of learners’ vocabulary learning. For instance, Chen and Hsu (2020) investigated 66 EFL college students’ RV by means of a serious game. They divided target words into high, medium, and low-frequency groups and then found that students learned more high-frequency words (occurring more than six times) than the mediate and low-frequency words; nevertheless, it is still possible for students to learn lower-frequency words if they occur in a meaningful context. Laufer and Rozovski-Roitblat (2011) claim that repeated exposure to words helps learners enhance their memory. Chen and Hsu (2020) provide further evidence that repetitive exposure to words has a positive influence on incremental vocabulary learning within the gaming context. Thus, it seems that providing more encounters with target vocabulary to learners for an extended period of time may serve as a considerable factor in practical game-based teaching and learning.

DGBVL for Productive Vocabulary Learning

Few studies have explored the relationship between DGBVL and PV, and of these, findings show mixed results. For example, deHaan et al. (2010) conducted a 20-minute music video game intervention experiment and found that game-players recalled significantly less PV than game-observers owing to the extraneous cognitive load induced by the gameplay. In a similar vein, Franciosi (2017) examined participants’ knowledge of productive words. In the quasi-experiment, participants in both experimental and control groups were assigned Quizlet drill exercises as their homework after class. During regular 60-minute class sessions, the experimental group used the computer game Energy City, while the control group did not. After the session, participants completed a writing task and were asked to write as much as possible, using 33 target words chosen from the game Energy City. Both curricular and extracurricular settings were
found to be moderately effective. However, in a follow-up longitudinal study, only a small positive correlation ($r = 0.29$) was reported between the 6-week gameplay amount and the occurrences of target words in the writing task. No apparent influence was found between mastery of vocabulary in Quizlet and target words used in the writing. In short, the few studies that focus solely on PV have not provided robust evidence of DGBVL’s effectiveness for PV gains.

**DGBVL for Both Receptive and Productive Vocabulary Learning**

There is a notable paucity of studies that investigate the influence of DGBVL on both PV and RV. Sundqvist and her co-researchers have conducted a series of studies in recent years (e.g., Sundqvist, 2019; Sundqvist & Wikström, 2015; Sylvén & Sundqvist, 2012) to evaluate digital games’ impacts on students’ English RV and PV acquisition in extramural environments. Their research suggests that a positive relationship exists between digital gameplay and vocabulary. They found that gameplay time could serve as an important predictor of L2 vocabulary, but game type only had a mediate effect within the COTS games they employed. Accordingly, they suggest learners invest more time in gameplay to acquire more incidental vocabulary and recommend further research into game type’s interplay effects on vocabulary learning. However, some extraneous variables (e.g., participants’ unequal gameplay time) might affect the tenability of the research results discussed above.

More recently, some scholars have begun to consider the multidimensionality of word knowledge and to explore DGBVL efficacy from a lexical perspective. Rasti-Behbahani and Shahbazi (2020) randomly assigned 124 Iranian EFL learners to experimental (60-minutes of pair-work gameplay followed by a vocabulary test with access to an online dictionary) and control groups (pair-work on the vocabulary test with access to an online dictionary). Both groups performed equally well in both receptive and productive tests and the DGBVL task effectively enhanced the experimental group participants’ acquisition of productive-recognition of form-meaning. Although the researchers asserted that DGBVL could enhance the acquisition of receptive, productive, recognition, and recall knowledge of orthography, meaning, and association, several issues with their research design weaken their conclusion’s validity. First and most glaring, the reliability and validity of self-designed vocabulary tests in this study were not reported; second, the homogeneity of the two groups participants’ initial English proficiency was not analyzed before the treatment. In addition, the treatment time was only 60-minutes; thus, the long-term effects of DGBVL on both RV and PV remain unclear. In brief, although previous research has investigated PV and RV, these studies have not provided robust evidence for the efficacy of DGBVL on PV and RV, particularly in the context of long-term learning in the language classroom.

To summarize, while recent studies have examined DGBVL from different angles, several potential issues are worth noting. First, many studies do not clearly define their vocabulary or word knowledge type, instead opting to generalize using the umbrella terms of “vocabulary learning” or “vocabulary acquisition” (e.g., Ranalli, 2008; Wu, 2018). This makes it difficult to distinguish what exact word knowledge the studies investigated (e.g., RV or PV). Second, in general, most existing studies focus on short-term incidental or intentional vocabulary learning through DGBL. However, as claimed by some researchers (Cobb & Horst, 2011), it appears that a longer period of gameplay under direct instruction (i.e., explicit teaching of specific skills) may be necessary to consolidate vocabulary learning. Third, the generalizability of some published studies on DGBVL’s overall efficacy is problematic, especially concerning its retention effects; this is because effect sizes are frequently not reported, and the treatment durations and the spacings between posttests and delayed posttests vary across different studies. Therefore, it remains to be seen what long-term effects DGBVL may have on RV and PV learning, especially in the context of direct instruction within the EFL classroom. Accordingly, this study aimed to address the following research questions:

1. Does DGBVL have a significant influence on EFL learners’ long-term RV learning?
2. Does DGBVL have a significant influence on EFL learners’ long-term PV learning?
Methods

These research questions were explored through an 18-week longitudinal mixed repeated measures quasi-experiment design.

Participants

Participants were recruited by means of convenience sampling from two intact freshmen EFL classes of a key university in Eastern China; 23 (4 males and 19 females) were in the experimental group (EG) and 25 (1 male and 24 females) were in the control group (CG). Such low ratios of male to female students are common for many English major classes in China universities (see the large-scale stratified survey in You and Dörnyei, 2016). All participants passed CET-4 (College English Test Band 4, a national English proficiency test for undergraduates and graduates in China), but had not reached CET-6 (the higher-level test in this national testing system), indicating their intermediate English proficiency. Their ages ranged from 18 to 19. To avoid the heterogeneity that might be brought by convenience sampling, outliers in pretest scores were examined first and an independent samples t-test was used later to compare the means of the two classes. Participants in both groups had no significant differences in their pre-tested RV scores ($M_{EG} = 12.26, M_{CG} = 10.92, t_{(46)} = 1.654, P_{RV} = 0.105 > 0.05$), and PV scores ($M_{EG} = 9.96, M_{CG} = 9.92, t_{(46)} = 0.38, P_{PV} = 0.970 > 0.05$). All students voluntarily participated in the experiment and could drop out at any time. They understood that the pre and post vocabulary test scores had no impact on their daily or final exam scores in the semester.

Materials

**Digital Games: Match, Gravity, and Live on Quizlet App**

The *Match*, *Gravity*, and *Live* games in the *Quizlet* App were utilized for the EG. These vocabulary games’ learning activities were automatically generated by the *Quizlet* system based on study sets that were created from the vocabularies of each unit in the *Comprehensive English Coursebook* (see more details in the following section).

**Figure 1**

*Solo Games Match (Top) and Gravity (Bottom)*
**Match** (see Figure 1) is a single-player game that aims to strengthen and examine players’ RV knowledge. It asks players to match words with definitions. For each round, once all words are correctly paired together, the players receive feedback, including their completion time, rank, and some additional comments regarding their performance (e.g., “That was fast!”, “Congratulations!”). Players can also send invitations to their friends to challenge their best records via social media (e.g., WeChat).

Similar to **Match**, **Gravity** (Figure 1) is also a single-player game, but it focuses on both receptive and productive vocabularies. The target vocabulary is embedded in asteroids that fall from the sky. As they fall, players must correctly spell out the words or type out their correct definitions. If players miss a word twice, a red asteroid will destroy players’ home planet. Players can choose between answer modes (English, Chinese, or random), difficulty level (easy, medium, or hard, each with increasing falling speed), and word bank (all words in the study set or self-chosen words). At the end of the game, feedback is provided to players, such as their scores, leadership rank, and award badges for good performance. Players with top scores can then proceed to the next level.

**Live** (Figure 2) is a collaborative multi-player game that solely addresses RV. Players work in small teams (3–4) to learn a Quizlet study set by finding the word that matches a particular definition, while racing against the other teams in the class. Since no individual team member has all of the answers, team members must cooperate and communicate in order to win; this insures engagement and contribution from each team member. **Live** facilitates more accuracy than speed because once a member chooses an incorrect answer, the whole team must start over from the beginning. Because of these game features, both students’ vocabulary proficiency and communication are likely to be reinforced.

**Figure 2**

**Live: A Cooperative Game**

---

**Lexical Profiles in Comprehensive English Coursebook**

*Comprehensive English Coursebook* is a theme-organized college-level English textbook containing 12 units. Each unit involves about 30 target words (totaling 386 words) that occur one to two times in an associated article. Target words are listed in page margins with their phonograms, Chinese equivalents, English definitions, and corresponding proficiency level (CET-4 or CET-6). Each unit’s target words were input and edited by the course instructor into a Quizlet study set that could be automatically transferred into **Match**, **Gravity**, and **Live** games by means of the Quizlet application.
Vocabulary Tests

A vocabulary pretest and posttest were utilized to examine participants’ vocabulary learning achievements. While identical in items, the order of items was shuffled in order to avoid practice effect. Both the pretest and posttest comprised two parts (see Appendix). Considering the need to go beyond participants’ CET-4 English proficiency level, two to three CET-6 level target words were randomly selected from each unit. This generated a total of 40 target words which were then divided equally into two parts. Part I and II aim to test participants’ RV and PV knowledge, respectively.

Part I was in the format of meaning recall; every target word was displayed with underlined and bold font in a contextual sentence chosen from Collins English-Chinese Dictionary. Participants were asked to write down the corresponding Chinese equivalents. Part II adopted a cued recall format, where a L2 word’s first letter was primed in a contextual sentence and the L1 equivalent was presented to help participants spell out the exact word for testing rather than an alternative (Lindstromberg, 2020). Two items from each part are provided as examples:

Example from Part I:  *He spent his adolescent years playing guitar in the church band.*
Example from Part II:  *Can you c____________ music on the computer? (创作)*

The content validity of the vocabulary tests was established by two experienced experts who are specialized in language testing. They verified that the test items were congruent with target words and that all test items were clear enough to be answered from the sentences presented.

Marking criteria were developed to optimize reliability and objectivity. For RV in Part I, variable Chinese answers for each test item were discussed, with several acceptable synonyms being agreed upon. For productive vocabularies in Part II, since “constructed-response items are inherently less controlled, and may be more difficult to score, than selected-response items” (Read, 2020, p. 550), standardized guidelines for marking were created to ensure inter-rater reliability and to take into account the differences in the spelling and grammatical form of target words (e.g., singular or plural noun forms). For spelling, any missing, added, wrong letter, or wrong order per word was regarded as a wrong answer. For grammar, errors in suffix endings involving -ed, -s, and -ing were ignored. The purpose of these criteria was to focus on accurate vocabulary knowledge while grading objectively with acceptable variations in meanings and forms (Hirsh, 2015). Following these criteria, the correct Chinese interpretation and English spelling for each item in the pretest and posttest were graded for 1 point, with a total score of 40 points possible for each test.

After being trained in the above-mentioned marking criteria, two experienced instructors graded each test paper independently. The inter-rater reliability of marking accuracy was excellent, with intra-class correlation coefficients (ICC) being 0.998 with 95% Confidence Intervals (CI) [0.997, 0.999] for the RV pretest, and 0.992 with 95% CI [0.991, 0.996] for the RV posttest. As for the PV pre- and posttests, the ICC was 0.998 with 95% CI [0.996, 0.999] and 0.998 with 95% CI [0.997, 0.999] respectively.

Subsequently, any disagreements between the two raters were discussed and a consensus was ultimately reached on pretest and posttest scores. Afterward, the Kuder-Richardson formula 20 (KR-20) was leveraged to establish the reliability of the tests since the items in both tests were dichotomous (Woodrow, 2014). The KR-20 reliability of the pretest and posttest were 0.793 and 0.916 respectively, signifying that the vocabulary tests were indeed reliable and suitable for further analysis.

Research Procedures

Two groups had their Comprehensive English classes on the same day of each week during the entire semester. They had the same Chinese instructor, who had 15 years of experience teaching EFL to University English majors. The EG adopted the game-based learning instruction approach to learn target words by using the game functions embedded in the Quizlet App (as described in the Materials section) on their laptops or smartphones during classes. The CG received instruction by means of a teacher-centered PowerPoint lecturing method to learn target words. The instructor explained the meanings, usages, and sentence examples of target words with the assistance of PowerPoint slides, and participants finished the
vocabulary match and fill-in-the-blank exercises in their textbooks in class. Except for the different teaching approaches, other factors, such as the learning content and pace of each class, were kept the same in order to minimize potential influences from extraneous variables.

The overall experiment lasted for 18 weeks and consisted of preparation, treatment, and post-experiment phases (see Figure 3).

**Figure 3**

*Experiment Flowchart*

In the preparatory phase, the experiment’s purposes, procedures, and irrelevance to their formal academic achievements were explained to the students in both groups. Consent forms were then collected from all participants. The vocabulary pretest was also completed by students in the first week under the instructor’s supervision. In Week 2, the EG received a 45-minute Quizlet App utilization training which included instruction on how to install the App, as well as game tutorials and guided gameplay with non-target words.

The treatment phase lasted from Week 3 to 18. Each week, both groups spent 30 minutes learning target words in the *Comprehensive English Coursebook*. Students in the EG played *Match*, *Gravity*, and *Live* games, while those in the CG learned through teacher-instructed PowerPoint lecturing and textbook exercises.

The immediate vocabulary posttest was conducted during a 45-minute class in Week 18, under the instructor’s supervision and without reference materials.

**Data Analysis**

**Data Analysis Methods**

A repeated-measures analysis of variance (RM-ANOVA) was conducted by using SPSS 26, following a General Linear Model with group (i.e., EG and CG) as a between-subjects variable, the time of tests (T1 pretest and T2 posttest) as a within-subjects variable, and the RV and PV test scores as dependent variables. In order to better distinguish between the effects of these independent variables on vocabulary learning, a
RM-ANOVA was selected as the means of statistical analysis because it allows investigating interaction effect between within-subjects and between-subjects factors and is very sensitive in detecting even the slightest variations in the levels of within-subjects factor (Verma, 2015). It is also statistically powerful as it factors out some within-group subject variation and can be used with smaller sample sizes (20 or less) (see Larson-Hall, 2015).

**RM-ANOVA Assumptions Tests**

Missing data, outliers, and possible statistical assumption violations were examined with SPSS Frequencies, Explore, and One-Way ANOVA procedures before commencing with further data analysis. No missing values were found in the pre- and posttest scores. Several univariate outliers were detected, but they were deemed to be a small part of the population under study (i.e., three top and bottom students) and not extreme or unusual enough to require deletion (Gamst et al., 2008). Skewness and kurtosis values of each pre- and post- vocabulary test scores (Table 1) were between ± 3.0 and ± 7.0, indicating normal distribution was still within an acceptable range, although the Shapiro-Wilk test results of some groups’ scores were statistically significant (p < 0.05) (Gamst et al., 2008). The Bartlett and Brown–Forsythe tests of the equality of error variances assumption were statistically significant for posttest scores (RV and PV) (p < .05) but not for pretest scores (RV and PV), indicating homogeneity across independent variable groups in pretests scores but heterogeneity in posttests scores (Table 1). This heterogeneity might be attributed to the small sample size (Gamst et al., 2008). Lastly, because the variable in this experimental study had only two levels, the data is not suitable to be tested for sphericity and does not need to follow the sphericity assumption, which only applies when there are more than two comparisons to be made in a variable or interaction term (Larson-Hall, 2015). A Greenhouse–Geisser correction was referred to in the following result reports. Given the literature that argues that the ANOVA test is robust against the violation of the normal assumption (Schmider et al., 2010), and that even if the homoscedasticity assumption is violated, results may not be necessarily invalid (Larson-Hall, 2015; Brezina, 2018), we decided to proceed with the RM-ANOVA operation.

Two independent samples t-tests were conducted for between-subjects (group) and within-subjects (time) in order to investigate potential differences between treatment groups and to make further comparisons.

**Results**

The descriptive statistics for the two groups’ RV and PV scores for T1 and T2, as well as the results of the Shapiro-Wilk test and Test of Homogeneity of Variances are reported in Table 1. Mean values indicate that participants in both groups achieved solid gains in RV and PV after one-semester of treatment; this is made apparent from the much higher T2-RV and T2-PV scores in each group. In addition, the mean values of T2-RV and T2-PV in the EG exceed those in the CG. The statistical significance of these differences for each test is further explored in the following sections.

**Findings Related to RQ1: DGBVL’s Influence on EFL Learners’ RV**

The first research question, concerning whether or not DGBVL has a significant influence on EFL learners’ long-term RV learning, was positively confirmed in several ways. First, the results of the mixed 2×2 RM-ANOVA analysis on the RV scores (Table 2) reveal a significant interaction effect between group and time ($F_{(1/46)} = 44.442, p = .000, \eta^2_p = 0.491$). This indicates that growth in RV from T1 to T2 was significantly different between the two groups, and that the effectiveness of the two instructional approaches was not the same (as illustrated by the steeper slope of EG in Figure 4). The EG improved significantly more (RV$_{T2}$ – RV$_{T1}$ = 7.39) than the CG (RV$_{T2}$ – RV$_{T1}$ = 2.84), as shown in Table 3. The effect size of the difference, expressed in partial eta square ($\eta^2_p$), shows that this two-way interaction accounts for 49.1% of the variance in RV scores, which represents a large effect size (Cohen, 1988).

Second, the results of the RM-ANOVA analysis on RV scores (Table 2) indicate that there was a significant main effect of the independent variable of treatment group ($F_{(1/46)} = 32.058, p = .000, \eta^2_p = 0.411$). This
signifies that there are significantly different effects between the DGBVL and PPT lecturing approaches, although this influence from the treatment group variable ($\eta_p^2 = 0.411$) is less than the main effect of the variable of time ($\eta_p^2 = 0.825$).

**Table 2**

*Mixed RM-ANOVA Results on RV*

<table>
<thead>
<tr>
<th>Source</th>
<th>Type II Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared ($\eta_p^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group</td>
<td>313.357</td>
<td>1</td>
<td>313.357</td>
<td>32.058</td>
<td>0.000</td>
<td>0.411</td>
</tr>
<tr>
<td>Error</td>
<td>449.633</td>
<td>46</td>
<td>9.775</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>Time</td>
<td>605.010</td>
<td>1.000</td>
<td>605.010</td>
<td>216.716</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Time*Group</td>
<td>124.070</td>
<td>1.000</td>
<td>124.070</td>
<td>44.442</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Error</td>
<td>128.419</td>
<td>46.000</td>
<td>2.792</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3**

*Paired Samples t-Test Analysis on Within-Subjects (Time)*

<table>
<thead>
<tr>
<th>GROUP</th>
<th>Test</th>
<th>Time</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
<th>df</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EG (N=23)</td>
<td>RV</td>
<td>1</td>
<td>12.26</td>
<td>2.96</td>
<td>12.55</td>
<td>22</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>19.65</td>
<td>0.65</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PV</td>
<td>1</td>
<td>9.96</td>
<td>3.28</td>
<td>15.43</td>
<td>22</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>19.35</td>
<td>1.27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG (N=25)</td>
<td>RV</td>
<td>1</td>
<td>10.92</td>
<td>2.66</td>
<td>7.71</td>
<td>24</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>13.76</td>
<td>2.93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PV</td>
<td>1</td>
<td>9.92</td>
<td>3.32</td>
<td>7.80</td>
<td>24</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>13.32</td>
<td>3.08</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* T1 = Pretest, T2 = Posttest, RV = Receptive Vocabulary, PV = Productive Vocabulary, EG = Experimental Group, CG = Control Group
**Figure 4**

*RV and PV Scores Across Testing Time in Different Groups*

![Graphs showing estimated marginal means of RV and PV scores across testing time in different groups.](image)

**Table 4**

Independent Samples *t*-Test Analysis on Between-Subjects (Group)

<table>
<thead>
<tr>
<th>Test</th>
<th>Time</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
<th>df</th>
<th>Sig.(2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RV</td>
<td>1</td>
<td>EG</td>
<td>23</td>
<td>12.26</td>
<td>2.96</td>
<td>1.65</td>
<td>46</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CG</td>
<td>25</td>
<td>10.92</td>
<td>2.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>EG</td>
<td>23</td>
<td>19.65</td>
<td>0.65</td>
<td>9.79</td>
<td>26.53</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CG</td>
<td>25</td>
<td>13.76</td>
<td>2.93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PV</td>
<td>1</td>
<td>EG</td>
<td>23</td>
<td>9.96</td>
<td>3.28</td>
<td>0.04</td>
<td>46</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CG</td>
<td>25</td>
<td>9.92</td>
<td>3.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>EG</td>
<td>23</td>
<td>19.35</td>
<td>1.27</td>
<td>9.00</td>
<td>32.43</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CG</td>
<td>25</td>
<td>13.32</td>
<td>3.08</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. T1 = Pretest, T2 = Posttest, RV = Receptive Vocabulary, PV = Productive Vocabulary, EG = Experimental Group, CG = Control Group

To investigate whether the treatment group difference existed either before or after the experiment, an independent samples *t*-test was conducted for the between-subjects (group) to make further comparisons. The results in Table 4 reveal that before the experiment treatment, the EG’s RV proficiency (\(M = 12.26, SD = 2.96\)) was slightly higher than that in the CG (\(M = 10.92, SD = 2.66\), however, there was no significant difference (\(t_{(46)} = 1.65, p = 0.11\)). In contrast, the two groups’ posttest outcomes in RV showed that the EG (\(M = 19.65, SD = 0.65\)) significantly outperformed the CG (\(M = 13.76, SD = 2.93\)), with \(t_{(26.53)} = 9.79, p = 0.00\). These results suggest that the DGBVL instruction was more effective in improving participants’ RV knowledge than the traditional PPT lecturing approach. Furthermore, the standard deviation (SD) in EG changed from 2.96 to 0.65 after the experiment treatment. However, the SD values for the CG before and after the experiment remained nearly the same (\(SD_{T1} = 2.66, SD_{T2} = 2.93\)). This likely indicates that DGBVL was instrumental in narrowing down the disparity in vocabulary knowledge among the participants.

In short, in regard to RQ1, the data suggests that both time and teaching strategy had a significant effect on participants’ achievements of RV knowledge; however, time’s influence outweighs teaching approaches’ power.

**Findings Related to RQ2: DGBVL’s Influence on EFL Learners’ PV**

Statistical results from the mixed RM-ANOVA analysis of treatment group, time, and PV (Table 5)
confirmed that DGBVL had a significant influence on the EFL learners’ PV learning results. First, a significant interaction effect was found between time and treatment groups ($F_{(1/46)} = 65.610, p = .000, \eta_p^2 = 0.588$); this indicates that the PV improvements under the 16-week of instruction between the two groups are significantly different (also illustrated in Figure 4). Furthermore, the effects from the interaction are rather large, as indicated from the partial eta square value ($\eta_p^2 = 0.588$), explaining 58.8% of the variance in PV scores. The finding that participants in the EG gained more significant growth in PV knowledge than those in the CG is also strengthened by the mean difference between post- and pretest PV scores (EG: $PV_{T2} – PV_{T1} = 9.39$, CG: $PV_{T2} – PV_{T1} = 3.4$, see Table 3).

Time had a significant main effect on PV learning, as made evident from the mixed RM-ANOVA test results for PV (see Table 5, $F_{(1/46)} = 288.001, p = .000, \eta_p^2 = 0.862$); this effect size is rather large. The paired sample t-test (Table 3) indicated that both groups improved their PV knowledge after 16-week longitudinal treatment under both teaching approaches. Specifically, the EG achieved significantly higher PV scores in the posttest ($M = 19.35, SD = 1.27$) than that in the pretest ($M = 9.96, SD = 3.28$), with $t_{(22)} = 15.43, p < 0.05$. Also, the CG’s gains in PV knowledge were statistically significant, although not as large as EG’s gains ($M_{T2} = 13.32, SD_{T2} = 3.08, M_{T1} = 9.92, SD_{T1} = 3.32$), with $t_{(24)} = 7.80, p < 0.05$.

Table 5

<table>
<thead>
<tr>
<th>Source</th>
<th>Type II Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-subjects Group</td>
<td>220.275</td>
<td>1</td>
<td>220.275</td>
<td>16.537</td>
<td>.000</td>
<td>0.264</td>
</tr>
<tr>
<td>Error</td>
<td>612.715</td>
<td>46</td>
<td>13.320</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-subjects Source</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>943.760</td>
<td>1.00</td>
<td>943.760</td>
<td>288.001</td>
<td>.000</td>
<td>0.862</td>
</tr>
<tr>
<td>Time*Group</td>
<td>215.000</td>
<td>1.00</td>
<td>215.000</td>
<td>65.610</td>
<td>.000</td>
<td>0.588</td>
</tr>
<tr>
<td>Error</td>
<td>150.739</td>
<td>46.00</td>
<td>3.277</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results from the RM-ANOVA test, shown in Table 5 also indicate the significant main effect of the treatment group, that is, the efficacy of DGBVL instruction is significantly different from PPT lecturing instruction ($F_{(1/46)} = 16.537, p = .000, \eta_p^2 = 0.264$). Results of the independent samples t-test, listed in Table 4, indicate that before treatment the EG group ($M_{T1} = 9.96, SD_{T1} = 3.28$) had roughly the same PV proficiency as the CG group ($M_{T1} = 9.92, SD_{T1} = 3.32$), with $t_{(46)} = 0.04, p = 0.97$. After treatment, however, significantly different PV proficiency scores were found between the two groups, with the DGBVL group’s posttest scores ($M_{T2} = 19.35, SD_{T2} = 1.27$) much higher than those in the PPT group ($M_{T2} = 13.32, SD_{T2} = 3.08$), with $t_{(32.43)} = 9.00, p = 0.00$ (correction values are reported here since equal variance was not assumed in PV posttest scores). Similar to RV, a decrease in the SD was also observed for the EG, but not in CG, in PV scores. Thus, it can be concluded that DGBVL appears to be more effective than PPT lecturing for participants’ mastery of RV and PV.
Discussion

This study sought to determine DGBVL’s long-term effects on both EFL learners’ RV and PV learning. The findings revealed that DGBVL was efficacious for both RV and PV knowledge learning. It was also found that time, or prolonged use of DGBVL in EFL classroom instruction, played a vital role in students’ RV and PV learning process.

The findings that treatment time and instruction types have significant interaction effects echo previous studies investigating DGBVL’s performance in RV and add new evidence of its effects on PV. First, in relation to those studies that solely investigated RV, this study further supports the claim that gamified vocabulary applications can lead to successful vocabulary gains (Dindar et al., 2021), even more so than traditional learning approaches (e.g., teaching via PowerPoint) (Wu, 2018). Second, in light of the literature focusing solely on PV, the DGBVL’s effectiveness found in this study is more encouraging. For instance, Franciosi (2017) found that PV score differences between the EG and CG had a medium-size effect ($\omega^2 = 0.1$), and that a small positive correlation existed between gameplay time and PV occurrences in a follow up writing task. In the current study, DGBVL was found to be more effective for PV learning, with a large size effect ($\eta^2 = 0.264$). This larger effect size might be the result of the different research designs leveraged.

In Franciosi’s (2017) study, 60-min DGBL intervention was added to EG as an extra enhancement after participants’ extracurricular Quizlet vocabulary drill exercises. In the current study, however, Quizlet games were integrated into the EG’s 16-week curricular learning process as the primary instructional approach. In addition, while Franciosi tested participants’ PV performance by counting the target words in the writing task, this study measured the productive form-meaning recall of target words. Third, the present study’s results also corroborate Rasti-Behbahani and Shahbazi’s (2020) findings, who found DGBVL to be a more effective treatment for both RV and PV learning.

Another finding worth discussing is that the main effect of time outweighed one of the instructional approaches, at least in its explanatory power of RV and PV gains, which suggests that time is a factor that should not be ignored. Previous research has demonstrated a significant positive correlation between gameplay time and RV and PV incidental learning in the long-run extracurricular context (e.g., Sundqvist, 2019; Sundqvist & Wikström, 2015). This 16-week study adds further evidence that long-term exposure to DGBVL in the formal context with direct instruction and intentional learning can also have a significant positive effect on RV and PV gains, and has promising potential in narrowing down the disparity in vocabulary knowledge amongst learners. A possible explanation for this is that longer treatment time is more likely to guarantee multiple and repetitive exposures to target words in the gaming context, which consequently results in a positive influence on incremental learning of vocabulary knowledge (Chen & Hsu, 2020). Instead of repetitive and monotonous vocabulary drills, this study produced 36 (12×3) customized Quizlet games based on 12 units’ vocabulary study sets and 3 types of games (Match, Gravity, Live). Such games might generate interest or motivation for a broad range of individuals, rendering them more engaged in their studies over a 16-week long period, which is a commonly reported characteristic of DGBVL in previous studies (e.g., Calvo-Ferrer, 2017; Yang et al., 2020).

Furthermore, it is worth noting Spada’s (2005) observation that research conducted in real classrooms with real learners is more likely to inform classroom practice. The practice of importing a language coursebook’s vocabulary items directly into Quizlet and then incorporating the educational games into routine classroom practice is within the means of most language teachers around the world. Accordingly, since both formal and informal settings have initially witnessed the positive effects from DGBVL, practitioners and researchers would likely benefit from more empirical studies that identify the duration time of DGBVL instruction that is ideal for maximizing RV and PV gains, and explore the optimal combination of in-school and out-of-school DGBVL. While it is logical that small time frames may not offer enough exposure for long-term retention, it is also equally plausible that the efficacy of DGBVL instruction may eventually fade over long time periods as each exposure may eventually result in diminishing returns in motivation and engagement.
The relationship between gender and DGBVL is another issue worth investigating more fully. In some earlier studies, males were found more engaged in games and tended to dedicate more time to gameplay (Chou & Tsai, 2007). Sundqvist’s studies (Sundqvist & Wikström, 2015; Sylvén & Sundqvist, 2012) found a positive correlation between gameplay time and L2 proficiency and reported that boys outperformed girls in vocabulary learning. However, they argued that it was not gender per se that explained the found differences and they hypothesized that girls would equally benefit well in L2 proficiency if their gameplay time was similar to boys (Sylvén & Sundqvist, 2012). However, the above studies were based on extracurricular settings without equal gameplay time among participants. Other studies conducted in the context of language learning classes had different findings, such as Hitosugi et al. (2014), who found no significant effect for gender in three (pre-, post-, and delayed) vocabulary tests after five 50-minute L2 class sessions by using an off-the-shelf videogame. Similarly, no gender differences in DGBVL were found in Calvo-Ferrer and Belda-Medina’s (2021) study leveraging a 90-minute videogame or Hsu’s (2019) study on the use of 80-minute AR games in L2 classrooms. The current study provides further evidence that females also can benefit significantly from DGBVL in long-term formal settings with equal gameplay time.

Conclusion

This study aimed to better understand the potential value of DGBVL in helping EFL students to learn vocabulary. More specifically, it leveraged a quasi-experimental research design to investigate DGBVL’s long-term effects on RV and PV, within the educational context of the language classroom. The most important finding is that while participants in both experimental and control groups saw an increase in RV and PV by the end of 16 weeks of treatment, the vocabulary gains of the former group were significantly more than the latter.

The current study does, however, have some limitations. First, to assure face validity, only a limited selection of target words was used for form-meaning link testing; however, word knowledge is multifaceted and other types of word knowledge had not been examined to depict learners’ mastery of vocabulary more comprehensively in breadth and depth under longitudinal formal instruction. Second, the current study did not make use of a delayed posttest and therefore the retention rate of acquired RV and PV knowledge remains unknown. Third, the current study did not statistically compare individual differences among learners. There is some evidence, for example, that indicates that some factors, such as students’ favored game types, gender (Sundqvist & Wikström, 2015), and game frequency (Sundqvist, 2019) might produce various DGBVL effects, but the current study did not leverage such independent variables and compare corresponding differences.

Therefore, while this study provides evidence that language learners may benefit from long-term DGBVL within the classroom, we conclude that further research is needed in order to inform language practitioners as to how to use DGBVL most effectively. First, a fuller understanding is needed of how DGBVL-based instruction may assist with other types of word knowledge, such as word parts, associations, collocations, and the automaticity or speed of word processing. Second, the question as to what is the ideal amount of time for leveraging DGBVL in classroom instruction deserves further investigation; this includes identifying an ideal duration of time for intervention, the ideal frequency of exposure to target words, and the most suitable interval time between delayed and posttests. Lastly, it may be prudent for researchers to shift their attention from the effects of DGBVL on vocabulary learning to how learning context (e.g., formal or informal; incidental or intentional learning) and individual differences (e.g., gender, cognitive styles) can support or undermine the efficacy of using digital game-based language learning.

Acknowledgements

This research was supported by Jiangsu Education Science 13th Five-Year Plan Fund (Project Ref. No. C-b/2020/01/10), the Fundamental Research Funds for the Central Universities (Project Ref. No. SKYC2020027 and SKYZ2020023), the Humanities & Social Sciences Funds for the College of Foreign
Studies in Nanjing Agricultural University (WY202203), and Jiangsu Research Association of Educational Technology for Higher Institutions (Project Ref. No.2019JSETKT065). We would like to thank the reviewers for their meticulous review of the manuscript and many constructive and insightful suggestions.

Note

1. Table 1 is available at this following link:
   
   https://osf.io/wqg97/files/osfstorage/65fde68fae6a24004e1c6e1e

References


Appendix A. Vocabulary Test

Part I 请写出划线单词的汉语意思。
1. It's tedious work, but Mike seems to truly love it.
2. Showing the soles of shoes to someone is a sign of contempt in Arab culture.
3. Nobody will be cheerful on such a cheerless rainy day.
4. At Shanxi province, residents have said they almost choke on coal dust in the evenings.
5. The swimming pool is available only in summer.
6. The government monitored correspondence and telephone conversations.
7. The spectacular sunrise made us exclaim in surprise.
8. They set off in the car and bumped over a sandy track to Sam's farm.
9. Can you give me the highlight of your resume?
10. At the time, I was somewhat of a workaholic, clocking in sixty-hour weeks.
11. A study in 2002 showed that athletics help student academic performance in high school more than any other extracurricular activity.
12. Tom Chaney raised his rifle and shot him in the forehead, killing him instantly.
13. You can't do all the jobs yourself—you can delegate a task to your subordinate.
14. By the week's end, an extraordinary intimacy had grown up between Tom and his new friend.
15. Schedule a convenient time, at least once a week, to spend the whole day with your child.
16. I was disappointed by his indifference more than somewhat.
17. At the time, Iraq was seen as the biggest foe, followed by China and Iran.
18. He spent his adolescent years playing guitar in the church band.
19. The teacher drilled grammar and the multiplication tables every day.
20. The solar panel calculator is now as cheap as a pack of cigarettes.

Part II 请根据汉语释义和首字母提示拼写出该单词。
1. She's too p ______ to enjoy rude jokes! （古板的）
2. Can you c ______ music on the computer? (创作)
3. He has just come off a difficult a _________. （任务，作业）
4. They’ve decided to p ______ having a family for a while. （推迟）
5. I e _______ it will take three months to build the bridge .（估计）
6. This provides the extra motivation you need for those t ______ days. (困难的，艰难的)
7. It's been dry for so long that the forest could burst into f ______ at any moment. （火焰）
8. He greets me at the door with his signature boyish g ______ and a hug. （咧嘴笑）
9. The e__________ of bombs roused me out of a deep sleep. （爆炸）
10. To support himself economically, he did a lot of o ________ jobs, including shining shoes, washing dishes, and all that. (临时的，不固定的)
11. They were my friends, and I just didn’t want to e______ them publicly. (使尴尬，使困窘)
12. My secretary leaves us next week, so we are advertising for a r _________. （接替者）
13. A s______ of 100 winter-swimmers in different age groups indicates that 80 percent originally suffered from diseases of some kind. （调查）
14. He is a r__________ person to trust. （可信赖的，可靠的）
15. He is a **p**__________ writer, creating many great works in his life. (多产的)
16. The father told the son a **f**__________ about a puppet. (寓言)
17. They expected me to lie down like a **c**__________ . (懦夫)
18. Let us **p**__________ to next step. (继续进行)
19. These factories **c**__________ that they are not responsible for the pollution of the river. (声称)
20. The home is situated within easy **a**__________ of shops and other facilities. (接近，进入)
About the Authors

Wen Jia is an associate professor working at the College of Foreign Studies at Nanjing Agricultural University, China. She is also a Ph.D. candidate studying in the Department of Applied Linguistics at Xi’an Jiaotong-Liverpool University. Her research interests include second language acquisition, game-based learning, and technology-enhanced language learning.

E-mail: jiawen@njau.edu.cn

Liping Zhang is a professor working at the Department of Basic Courses in Army Engineering University of PLA. Her research interests include second language acquisition, English for Specific Purposes, and corpus linguistics.

E-mail: zlpworkgroup@163.com

Austin Pack is an assistant professor working at the Faculty of Education & Social Work at Brigham Young University-Hawaii. His research interests include the psychology of language learning, complex dynamic systems, network analysis, computer assisted language learning, and virtual reality technologies.

E-mail: austin.pack@byuh.edu

Yi Guan currently is an instructor at Global Institute of Software Technology in Suzhou, China. Her interests include mobile-assisted language learning and English vocabulary learning.

E-mail: guanyi95@sina.cn

Bin Zou is an associate professor at the Department of Applied Linguistics, Xi’an Jiaotong-Liverpool University, China. He received his Ph.D. in TESOL and computer technology from the University of Bristol, UK. He is the Founding Editor and Co-Editor-in-Chief of the International Journal of Computer-Assisted Language Learning and Teaching. Bin Zou is the corresponding author of this article.

E-mail: bin.zou@xjtlu.edu.cn