



Castledown

 OPEN ACCESS

Technology in Language Teaching & Learning

ISSN 2652-1687

<https://www.castledown.com/journals/tltl/>

Technology in Language Teaching & Learning, 6(2), 1–20 (2024)

<https://doi.org/10.29140/tltl.v6n2.1195>

The Effects of Gamified Daily Awards on Digital Vocabulary Flashcard Learning: A Case Study



LOUIS LAFLEUR 

Kwansei Gakuin University, Japan

louislafleur333@gmail.com

Abstract

Gamification in second language acquisition research is a popular topic, but there are few empirical comparative studies in the literature. This quasi-experimental mixed-approach study enrolled 77 Japanese university EFL learners to enable a comparison between two digital vocabulary flashcard learning software conditions carefully designed by the author; the gamified group's software (group 1; $n = 26$) had daily awards (i.e., consecutive day awards, medals related to daily participation, and a bonus point counter in an effort to encourage spaced learning principles and discourage cramming) and the non-gamified group's (group 0; $n = 51$) did not. The daily awards had a significant effect in encouraging the gamified group to spread out their study efforts throughout the 12-week study period more effectively than the control group as they completed a lower median number of tasks per active study day (non-gamified: 104.76, gamified: 82.11; $p = .021$, $r(75) = -.264$) but studied on more days to complete a similar total number of vocabulary tasks (non-gamified: 2313.00, gamified: 2228.50; $p = .601$). Moreover, the gamified group significantly outperformed the non-gamified group in terms of vocabulary knowledge score gains between the pretest and posttest; $p = .03$, $r(50) = .300$. These results show the numerous and significant positive pedagogical impacts of gamified daily awards.

Keywords: gamification, interleaved spaced repetition, vocabulary learning, case study

Introduction

Defining what a game is can be difficult; Reinhardt (2019) noted that just because educational activities are fun or gameful does not necessarily make them games. Moreover, there is no consensus on how

Copyright: © 2024 Louis Lafleur. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. **Data Availability Statement:** All relevant data are within this paper.

Table 1 *A Framework for the Use of Games in L2TL*

Type	Definition	Study Example
Game-enhanced L2TL	utilizes games that were not originally designed or designed for the purpose of L2TL. (e.g., commercial/off-the-shelf games)	“The Sims” Wang (2019)
Game-based L2TL	utilizes games that were specifically/intentionally designed for the purpose of L2TL. (e.g., educational games)	“Text Detective” Hugues (2023a & b)
Game-informed* L2TL	utilizes the theoretical principles of play and games (gamification) for L2TL purposes (e.g., gamified learning applications/activities)	“eigomemo.com” This study

Note: Table adapted from Reinhardt & Sykes (2014); *the term “game-informed” is often regarded as a synonym to the more popular term “gamification”.

many and/or what elements are required to constitute a game, and thus definitions remain quite vague such as Dempsey et al. (1996) who define a game as a set of activities that has constraints (i.e., rule-guided, and artificial in some regard), goals, consequences and incorporates some aspect of competition which can be with others or oneself. On the other hand, there is much less contention in defining types or categories of games. For Second Language Teaching and Learning (L2TL), the terms “game-enhanced”, “game-based”, and “game-informed” were proposed by Reinhardt and Sykes (2012, 2014) to delineate the use of games and gameful applications in L2TL (see Table 1).

According to Kapp (2017), the first documented print appearance of the term “gamification” in 2008 was followed by mainstream recognition by late 2010. In layman’s terms, gamification is the inclusion of game-design elements and principles in other fields to encourage and retain engagement. The influence of gamification spans through various fields such as the workplace (e.g., employee of the month-type awards to encourage workers to work more), and education (e.g., proficiency certificates/diplomas to encourage students to study more). According to Reinhardt (2019), the words “gamification”, “gamify” and “gamified” in the literature are often utilized/preferred when pointing to gameful elements that can be easily perceived (e.g., awards and score rankings); in contrast, the more recently introduced term “game-informed”, which concerns this paper’s area of research, is more likely to be used in contexts where the added gameful elements are not immediately perceived and seem to be more in line with general conventions and elicit less surprise (e.g., scheduled feedback). In fact, “game-informed” practices should not be viewed as a new concept as both current and past SLA theories and practices are to a certain extent already game-informed such as Task-based Language Teaching (TBLT) which aims to make language learning activities more meaningful and goal-oriented (Reinhardt, 2019).

Literature Review

Gamification in CALL

Since the advent of Computer-assisted Language Learning (CALL), the discussion and inclusion of game-informed elements/practices has been popular in academic circles (Reinhardt, 2019). As of the writing of this article, a quick search on Google Scholar reveals ~2,420,000 total article results for the term “computer assisted language learning” and ~627,000 results for this term combined with the term “game” (i.e., ~25.9% of total CALL articles). In terms of L2TL software, gamification implementations often relate to the inclusion of game-design elements such as earning points, completing challenges, and earning badges, certificates, awards, and/or principles such as aesthetics, and mechanics. The intended purpose of such implementations is to increase motivation towards learning and/or encourage autonomous and self-directed learning practices (i.e., decreasing learner reliance on formal resources

and instructors). Reinhardt (2019) identified smaller, limited games and educational apps that do not try to pass themselves off as full-fledged games but utilize some game mechanics as potentially holding new promise for language learning and teaching, and noted that it is this model that early developers of gameful CALL thought would be more likely to lead to long-term success as it leverages a learning-oriented disposition.

The Pitfalls and Benefits of Gamification

The literature stresses the importance of carefully implementing gamification features to avoid potential pitfalls (i.e., potential decrease in learners' motivation; Kapp, 2017; Reinhardt 2019). For example, the self-determination theory (Ryan and Deci, 2000) identified two possible reactions to recognition/feedback and extrinsic rewards. The first reaction lowers motivation for those who perceive them as controlling. The second reaction raises motivation for those who perceive them as being informative. Essentially, the self-determination theory implies that gamification elements (e.g., points, medals/rewards) should not be of a controlling nature (i.e., ensure voluntary participation) but encourage learners to build upon their autonomous learning skills. Autonomous learning skills are key to success in L2 learning according to various studies, for example, Sundqvist and Sylven (2016) identified early and sustained extramural L2 learning with overall L2 learning success. Another extensively researched benefit/potential of gamification is its ability to sustain learners' engagement. Mekler et al. (2017) found that gamification elements such as points and goals increased the total quantity of learning tasks completed by learners, especially for those who were autonomy-oriented, in comparison to the non-gamified control condition.

Game-Informed Language Learning and Vocabulary Acquisition

Although to the author's knowledge there are no papers that investigate the relationship between daily awards and vocabulary acquisition specifically, there are papers that investigate the impact of gamification features on vocabulary acquisition.

Calvo-Ferrer (2017) compared a non-gaming learning tool (i.e., control group) to an educational game (experimental group) and found that the experimental group significantly outperformed the control group in terms of learning outcomes. However, the paper should have implemented an Analysis of Covariance (ANCOVA; i.e., control for pretest scores), and not an Analysis of Variance (ANOVA; i.e., ignored the implications of the pretest scores) when calculating posttest learning outcomes, as this may very well have resulted in a learning outcome result over and not under the .05 significance mark.

Fithriani (2021) which compared a traditional learning method (i.e., drilling; control group) to an experimental group that used Quizlet (i.e., a gamified vocabulary learning web app) also found that the experimental group significantly outperformed the control group in terms of learning outcomes. In this study, the control group's scores barely increased between the pretest ($M = 6.48$) and posttest ($M = 6.5$) in comparison to the experimental group pretest ($M = 6.48$) and posttest scores ($M = 7.46$; $p = .00001$) which increased significantly. The control group's lack of progress during the treatment time frame is quite surprising and raises questions considering other studies which have used traditional learning methods (e.g., drills, wordlists, paper flashcards; Nakata, 2008; Elgort, 2011) have produced significant score increases between pre/posttests.

The Present Study

As noted previously, gameful learning has been shown to increase participant satisfaction/engagement and increase study times which has led to better learning outcomes in contrast to non-gameful/traditional

learning. However, there is a lack of studies that address other potential benefits of gamification such as the possibility of improving autonomous learning skills and study/learning efficiency via the gamified encouragement of spaced study practices (i.e., to discourage cramming and encourage spreading out one's study efforts) as pioneered by Ebbinghaus (1885/1964) and later systemized into spaced study systems by others such as Leitner (1972) and Lafleur (2015). This study aims to explore the acquisition of the latter section (items 501~963; 463 words) of the New Academic Word List (NAWL; Browne et al., 2013) through the comparison of two versions of digital vocabulary flashcard learning software carefully designed by the author to control the two research software conditions; one which includes gamified daily awards to encourage spaced learning and discourage cramming (i.e., consecutive day awards, medals related to daily participation, and a bonus point counter), and one which does not. In consideration of the various novel aspects and unknowns of this study, the following research questions were preferred to hypotheses:

RQ1: Do gamified daily awards have an effect on vocabulary software satisfaction?

RQ2: Do gamified daily awards have an effect on digital flashcard study habits?

RQ3: Do gamified daily awards have an effect on vocabulary learning outcomes?

Methodology

Summary

This quasi-experimental study uses a mixed-methods approach regarding participant data collection that relied heavily on quantitative data (test scores, software participation data, and Likert scale survey responses) but also included qualitative data (open-ended comments given within the post-study survey) to answer the posed research questions (see Table 2).

The Study Timeline

To assess learning outcomes, this study implemented the following research pattern (see Figure 1) over a total period of 12 weeks (i.e., week 1 start point; week 13 end point).

Table 2 *Summary of Methods and Instruments Utilized in this Study*

RQ# [Tool utilized]	Data collected	Data validity test(s)/Analysis method(s)	Descriptive statistics	Significance testing (p value)	Effect size calculation
RQ1 "Satisfaction" [Post-study survey]	Likert data, and Qualitative responses	Cronbach Alpha, and Braun & Clarke's (2006) thematic analysis approach	Mean, SD, CI, and % calculations	Mann-Whitney <i>U</i>	z-derived <i>r</i>
RQ2 "Study Habits" [Author's website app]	Participation data (e.g., task logs)	Shapiro-Wilk, Skewness/Kurtosis and z-scores data analysis	Median, IQR, min, max, and % calculations	Mann-Whitney <i>U</i>	z-derived <i>r</i>
RQ3 "Learning Outcomes" [pre/posttests]	Word knowledge data	Shapiro-Wilk, Skewness/Kurtosis and z-scores data analysis	Median, IQR, min, max, and % calculations	Mann-Whitney <i>U</i>	z-derived <i>r</i>

Note: SD = Standard Deviation; [CI] = 95% Confidence Interval, [Min, Max] = Minimum and Maximum values, z-derived *r* calculations followed Fritz et al. (2012) and Cohen's (1988) guidelines to calculate effect sizes for non-parametric data.

Context and Participants

A total of 77 volunteer Japanese university CEFR A2/B1 level learners of English majoring in business administration utilized the author's digital vocabulary learning software (see description below) under one of two software conditions (non-gamified; $n = 51$, and gamified; $n = 26$) over a period of twelve weeks and met the author's minimum requirements of 1000 tasks studied and demonstrated a score increase of at least one point between pre/posttests (if both were completed). A total of 52 participants (non-gamified; $n = 37$, and gamified; $n = 15$) took both the pre and posttest. Unequal group sizes were due to a convenience sampling approach being implemented (i.e., quasi-experimental study) to assure that participants from the same classroom had access to the same software condition (i.e., gamified or non-gamified) to avoid confusion among participants.

The Wordlist

The participants used one of the two versions of the author's software to gain study access to the final 463 frequency-ranked word items (items 501~963) of the New Academic Word List (NAWL; Browne et al., 2013) supplemented with contextualized sentences and translations (Kanazawa & Lafleur, 2023). The NAWL was selected because it is publicly available under a creative commons license, and most of its final 463 frequency-ranked word items were expected to be unknown by the participants which should aid in avoiding ceiling effects regarding pre/posttests results.

The Software

The wordlist was loaded into the author's digital flashcard learning software called "eigomemo.com". Eigomemo.com is a free web-based application accessible 24/7 via any web browser and any type of device (smartphone or tablet/laptop/desktop computer). Eigomemo.com, technically speaking, is Interleaved Spaced Repetition Software (ISRS) as it combines the concepts of spaced repetition (interval-based study) and task interleaving (the reoccurring practice or learning of multiple skills/concepts)

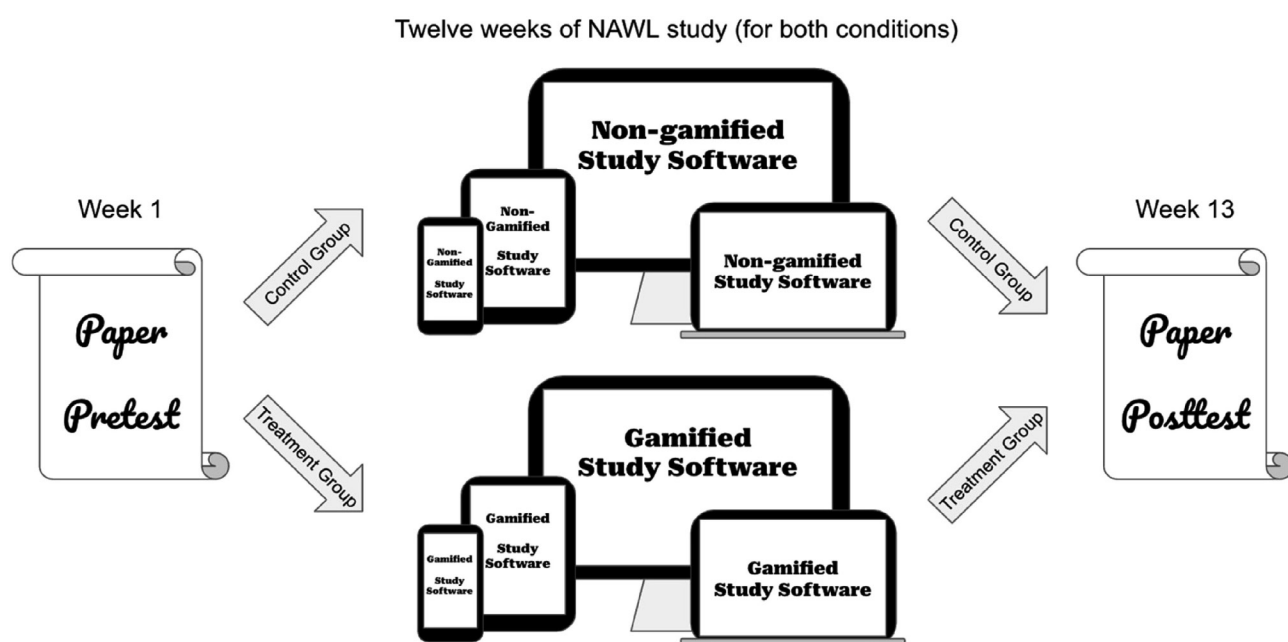


Figure 1 *The Effects of Gamified Daily Awards Study Timeline.*

in an effort/aim to promote both the long-term and intricate learning of words (Lafleur, 2020; see Appendix 1 & 2). In terms of vocabulary learning, task interleaving can provide a wider/deeper range of study, such as, focus on forms, four language skills, different task flows (L2 to L2, L2 to L1, L1 to L2), and both receptive and productive knowledge. ISR does not add extra tasks as additional items/flashcards but cycles between tasks according to the reached interval for every word item. In other words, the software customizes review/study for each user's specific learning needs (i.e., users who struggle with the spelling of particular word items will encounter these words' spelling tasks more often, and users who do not struggle with those words' spelling tasks will not encounter them as much). The task interleaving portion of the software was inspired by Nation's (2001) suggestion to implement an in-depth study of vocabulary to assure the correct use of words in terms of output and points to the importance of mastering all three features/elements of word knowledge: Meaning, Form and Use, and subsequent nine aspects of word knowledge which can be subsequently broken into receptive and productive areas of mastery.

Groups and Software Versions

All participants irrespective of their group (non-gamified or gamified) were given a study goal of 2400 or more tasks (i.e., completing one study task involves responding to one question related to one NAWL word item as determined by the software's algorithm, see Appendix 1 & 2) and were recommended a study pace of 200 tasks a week over the 12-week study period as this was judged as an appropriate amount of work to complete outside of class (i.e., approximately ~40 minutes per week). Moreover, all participants were given 10 minutes of class time per week (120 minutes total) to contribute towards reaching the set goal and were asked to complete the rest as homework/outside of class time. The "non-gamified" group's version software included a total progress counter (total number of completed tasks) and a recommended weekly goal progress counter. In addition to these counters, the "gamified" group's software version included pop-up medal animations, effects, and bonus points.

Three participation-linked bonuses (see Figure 2) inspired/informed by spaced learning principles (unlockable once a calendar day) would appear under the following set of conditions: (1) when the user completes all tasks/words items that were ready for review or study in a later interval, (2) when the user completes/adds ten new word items/flashcards to one's study mix after completing all tasks awaiting review/study, and (3) when the user studies for more than ten minutes (the software tallied how many seconds passed between the question and answer period of each task up to a maximum of 120 seconds, so breaks/inactivity would not significantly skew time records).

In no particular order, the first of these participation-linked bonuses to be completed would trigger a bronze medal animation, change the app's flashcard outline/contour to a bronze color for the remainder of the calendar day, and earn the user one bonus point. The second of these bonuses to be completed would trigger a silver medal animation, change the app's flashcard outline/contour to a silver color for the remainder of the calendar day, and earn the user an additional two bonus points. The third of these bonuses to be completed would trigger a gold medal animation, temporarily change the app's flashcard outline/contour to a gold color for the remainder of the calendar day and earn the user an additional two bonus points.

The consecutive day/bonus animation (see Figure 3) would be displayed after completing the first task on a given calendar day which followed a day or a number of days where the software was used consecutively for study. The user would be granted one bonus point if the software was used consecutively for two days, two points for three consecutive days, three points for four consecutive days, and so on. Finally, bonus points received from consecutive day awards and daily participation medals would be tallied on-screen within the bonus counter field (see Figure 3).



Figure 2 *Participation Linked Daily Awards/Points (Medals, Animations).*

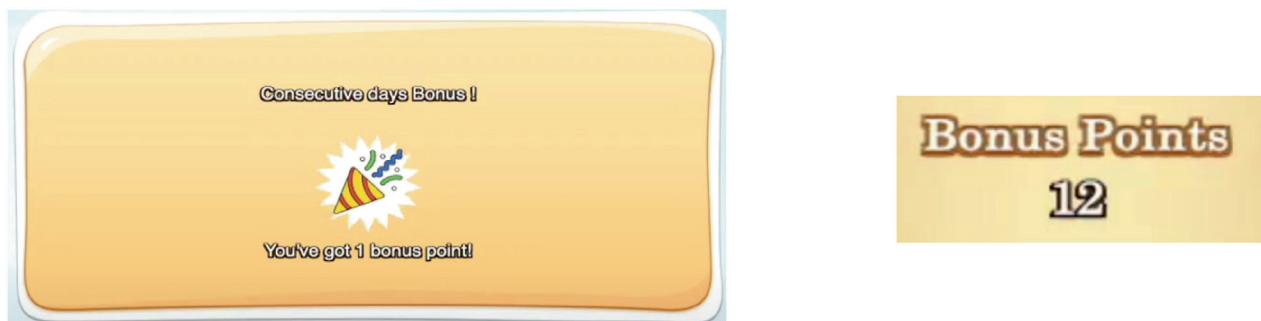


Figure 3 *Pop-up Consecutive Day Bonus/Animation, and Overall Bonus Point Counter.*

The Pre/Posttest (Pre/Post-Treatment Repeated Paper Test)

A recorded aurally driven 25-minute paper-based test was conducted twice, immediately before and after the twelve-week software treatment. An aural modality was preferred to a reading modality to avoid mistakenly giving away any of the “to-be written” answers to other questions, especially those regarding the same word item. The number of tested word items followed McLean, Stewart, & Batty’s (2020) recommendation to test at least 40 per 1000 words when assessing L2 form-recall and/or L2 meaning-recall modalities to permit reliable/acceptable inferences of total overall wordlist coverage/knowledge as this was sufficient to reach a Cronbach’s alpha value of .90 in their bootstrapping study. Following this logic (i.e., 40×463 total number of word items covered in this current study $\div 1000 = 18.52$), it was decided that at least 19-word items should be tested. Moreover, to control more reliably for item difficulty, word items were sequenced in their ordinal position of difficulty within each 100-word frequency band (i.e., subset) according to yes-no knowledge data collected in a previous study from Japanese university students (Kanazawa & Lafleur, 2023). From easier to more difficult, the 13th, 38th, 63rd, and 88th ranked word items of each 100-word band were selected to comprise the test (with the exception of the 88th word item of the final band not being selected since it only comprised a total

Table 3 *Pre/Posttest Sections and Task Flow*

Test sections	Task flow	Similar test format
(1) “Meaning” Listening Recall	After listening to a sentence and target word audio in English, the participant was asked to translate/write the target word in Japanese.	(McLean et al., 2021) Spoken Receptive Meaning-Recall/Listening meaning-recall
(2) “Form” Dictation Recall	After listening to the target word audio in English, the participant was asked to write its basic/dictionary form in English.	(Cheng & Matthews, 2018) Testing productive/phonological (ProPhon) vocabulary knowledge
(3) “Use” Listening Recall	After listening to the sentence audio in English of the target word, the participant was asked to write its translation in Japanese.	None, but inspired by Nation’s (2001) suggestion to enable a more “in-depth” learning/testing of vocabulary to ensure the correct “use” of the words.

of 63 words). This test was comprised of three sections for each of the selected word items (see Table 3 and Appendix 3), and were most similar to study tasks #1, #3, #5 in the study software (see Appendix 1 & 2):

In an effort to uphold good and fair results, the single grader (i.e., author) was blind to which of the two groups (i.e., gamified, and non-gamified) each of the 52 tests were associated with during the grading phase. The test was graded in the following manner: correct responses were given a score of one, incorrect responses were given a score of zero, and exceptionally some debatable/arguable responses were given a score of 0.5 (e.g., a sentence translation which omitted a single word/part that did not overtly compromise the sentence’s global meaning). Some aspects of the test were graded more leniently, in terms of word meaning, synonymous/viable answers were accepted, in terms of word form, part-of-speech variation of the same basic word form was accepted. On the other hand, some aspects of the test were stricter such as English spelling mistakes no matter how big or small were given a score of zero.

The Post-Study Survey

This post-study survey was administered with google forms to collect data pertaining to the effect of gamified daily awards on vocabulary software satisfaction (i.e., research question#1). As the survey was voluntary, it was completed by 60 out of the total 77 participants (i.e., 44 participants from the non-gamified group, and 16 from the gamified group).

The Analysis Procedures

For data reliability testing, Cronbach’s alpha was used to verify the internal consistency of Likert scale survey items. For the purpose of verifying the normality of distributions, Shapiro-Wilk, Skewness, and Kurtosis SPSS tests were conducted across this study’s various data points. These tests and calculations revealed some areas of concern, and further z score calculations uncovered some outliers in the data (e.g., one participant in the “non-gamified” group scored 3.93 standard deviations higher than their group’s average on the posttest). Instead of removing or reducing the effect of such data outliers, the author decided to use a Mann-Whitney U test, a non-parametric test, that uses a data point ranking system to circumvent abnormally distributed data for statistical significance testing.

The best way to conduct an effect size test for non-normally distributed data, according to Fritz et al. (2012), is to follow Cohen’s (1988) guidelines to calculate effect size by utilizing the z value: $r = z / \sqrt{N}$ (r effect size = z value divided by the square root of the sample number). This z value is commonly

reported as the “standardized test statistic” within SPSS when conducting a non-parametric test such as the Mann-Whitney *U*. In terms of descriptive statistics, with the exception of Likert scale data, median and interquartile range (IQR) calculations were preferred to mean and standard deviation (SD) calculations because of the presence of non-normally distributed data.

Regarding qualitative data, Braun and Clarke’s (2006) 6-phase guide to thematic analysis approach/framework was utilized to uncover, analyze and classify participant qualitative survey responses into a data extract (i.e., thematically coded chunk of data). For the purposes of this study, an essentialist/realist top-down theoretical approach to thematic analysis was conducted, and thus only comments related specifically to gamification were retained (i.e., focus was given to participants’ individual experiences/motivations that were related to the paper’s research theme).

Results

The collected data from the pre/posttests, author’s software, and post-study survey were used to create the following tables and figures, reveal elements of statistical significance and provide the basis for this paper’s discussion. Regarding the internal consistency of 5-point Likert scale survey items, a Cronbach’s alpha calculation revealed a score of .779 which is considered to be an acceptable level of reliability according to George & Mallery (2003). Information related to RQ1 can be found in Tables 4–6 and Figure 4, RQ2 can be found in Table 7, and RQ3 can be found in Tables 8 & 9.

Table 4 shows the survey data collected from 33 participants who had previous vocabulary software experience (52% of total respondents). The survey revealed the participants were most familiar with English Central, a digital 4-skills learning/practice environment to help learners build their vocabulary, and Quizlet, a learning app that incorporates flashcards and gamified learning tools. Other tools mentioned were Duolingo, and other lesser-known tools specifically designed for Japanese students such as “ターゲット [target]” and “mikan”. In regard to prior software overall satisfaction, no statistical difference was observed between both groups: Mann-Whitney $U = 113.000$, $p = .954$, $r(31) = -0.014$.

Table 5 shows the participants’ satisfaction rating of the author’s software under three categories (ease of use, interest, and usefulness). Overall, irrespective of group, interest scores were the lowest, and usefulness scores were the highest. This is not surprising as it matches the priorities of the original software design.

Table 4 *Participants’ Prior Vocabulary Study Software Experience and Satisfaction*

Group#	English Central	Quizlet	Duolingo	Other Software	Overall satisfaction Mean (SD) [CI]
Group (0) <i>n</i> = 23	10 (43%)	8 (35%)	4 (17%)	8 (35%)	3.43 (0.95) [3.03, 3.84]
Group (1) <i>n</i> = 10	7 (70%)	8 (80%)	1 (10%)	2 (20%)	3.40 (0.97) [2.71, 4.09]
Total (<i>n</i> = 33)	17 (52%)	16 (48%)	5 (15%)	10 (30%)	3.42 (0.94) [3.09, 3.76]

Note: (0) = Non-gamified Group, (1) = Gamified Group; 5-point Likert scale responses: 1 = very low satisfaction ~5 = very high satisfaction; (%) = % of participants, [CI] = 95% Confidence Interval.

Table 6 shows a comparison of participants' overall satisfaction of prior vocabulary software and their overall satisfaction of the author's software. It should be noted that these results are not a one-to-one comparison as this study's overall satisfaction software score is calculated from averaging ease of use, interest, and usefulness scores; see Table 4), in contrast to the prior vocabulary overall satisfaction score which was collected directly as its own survey question. Moreover, only 33 participants completed the satisfaction-related survey questions regarding both prior and treatment software. Although there were no results of statistical significance, it is interesting to note that the non-gamified group participants' satisfaction was slightly higher with software used previously, and this in contrast to the gamified group who were more satisfied with the study software than software they had used in the past.

Figure 4 shows all comments that pertained to gamified elements which were uncovered by implementing Braum and Clarke's (2006) thematic analysis approach. As the theoretical focus was on gamification, unsurprisingly uncovered comments were only provided by participants of the "gamified" group. Only 20% of the participants provided comments as it was optional when answering the survey. Of all the comments, only three were identified as being directly related to gamification. All three comments painted the gamified daily awards in a positive light and reflected the intended goals for which they were designed.

Table 5 *Participants' Study Treatment Software Satisfaction Rating*

Data point	Group#	Mean (SD)	[CI]	Mann-Whitney p value	Effect size z-derived r	%(0) ▽ (1) Mean
Software ease of use	(0) n = 44 (1) n = 16	3.45 (1.00) 3.56 (0.96)	[3.15, 3.76] [3.05, 4.08]	U = 371.000 p = .741	r = .043	▽ 50.00%
Software interest	(0) n = 44 (1) n = 16	2.84 (1.08) 3.19 (0.75)	[2.51, 3.17] [2.79, 3.59]	U = 427.000 p = .191	r = .169•	▽ 72.73%
Software usefulness	(0) n = 44 (1) n = 16	3.75 (0.94) 3.94 (0.68)	[3.46, 4.04] [3.58, 4.30]	U = 384.000 p = .569	r = .073	▽ 36.36%
Overall Combined score	(0) n = 44 (1) n = 16	3.35 (1.07) 3.56 (0.85)	[3.02, 3.67] [3.11, 4.01]	U = 404.500 p = .376	r = .114•	▽ 61.36%

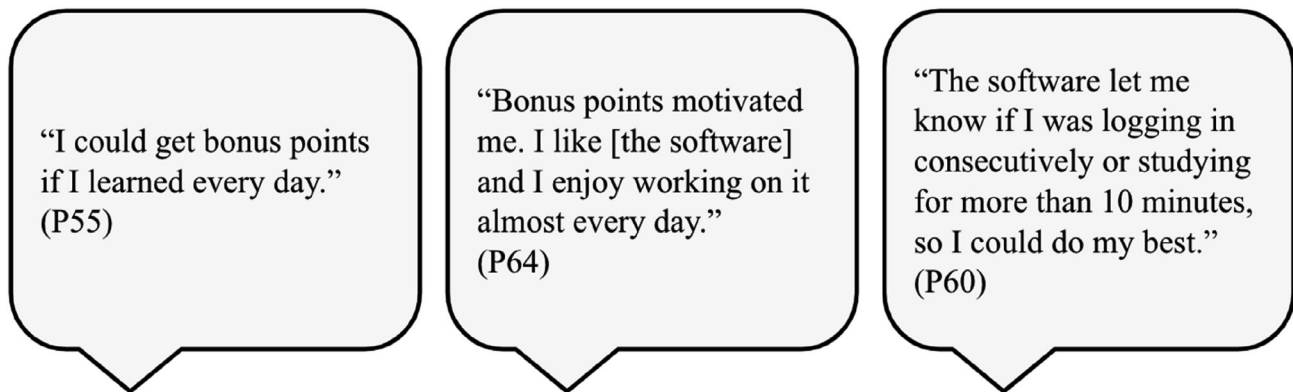
Note. (0) Non-gamified Group, (1) Gamified Group; SD = Standard Deviation, [CI] = 95% Confidence Interval (responses) 5-point Likert scale responses: 1= very low satisfaction ~5= very high satisfaction; %(0) ▽ (1) Mean = % of control group below experimental group's average; (Effect size) • = .100 ~ .300 Small effect size.

Table 6 *Prior and Treatment-Used Software Satisfaction (within groups)*

Group#	Prior Soft Satisfaction Mean (SD) [CI]	Study Soft Satisfaction Mean (SD) [CI]	Mann-Whitney p value	Effect size z-derived r
(0) n = 23	3.43 (0.95) [3.03, 3.84]	3.39 (0.92) [2.90, 3.88]	U = 258.000 p = .884	r = -.022
(1) n = 10	3.40 (0.97) [2.71, 4.09]	3.80 (0.59) [3.22, 4.38]	U = 65.000 p = .280	r = .261•

Note: (0) Non-gamified Group, (1) Gamified Group; (responses) 5-point Likert scale responses: 1 = very low satisfaction ~5 = very high satisfaction; (Effect size z-derived r) • = .100 ~ .300 Small effect size.

Table 7 shows the various participation-related data points collected from the participants using the two versions of the author's software. In many respects, both groups were quite similar and no differences of statistical significance were found in terms of tasks studied and total study minutes. The most important and statistically significant difference between the groups was the number of tasks completed per active day of study $U = 448.000$, $p = .021$, $r(75) = -.264$ which can be perceived as the gamified group embracing the spirit of spaced learning (see Figure 5). In other words, they spread out their study by completing less tasks per active study day on average but in the end completed a similar



Note. (P# = participant #), [the software] = .com author website used in the study.

Figure 4 Participants' Comments Related to Gamified Elements.

Table 7 Gamification Level Comparison and Software Participation Results

Data point	Group#	Median (IQR)	[Min, Max]	Mann-Whitney Test	Effect size z-derived r
# of tasks/questions completed	(0) $n = 51$ (1) $n = 26$	2313.00 (569) 2228.50 (745)	[1043, 2613] [1141, 2723]	$U = 614.500$ $p = .601$	$r = -.060$
# of active study days	(0) $n = 51$ (1) $n = 26$	20.00 (13) 24.50 (36)	[4, 52] [9, 77]	$U = 809.500$ $p = .114$	$r = .180$
# of tasks/questions per active study day	(0) $n = 51$ (1) $n = 26$	104.76 (73.14) 82.11 (87.39)	[45.27, 280.75] [33.42, 242.67]	$U = 448.000$ $p = .021^{**}$	$r = -.264$
# of total task/question study minutes	(0) $n = 51$ (1) $n = 26$	580.00 (229) 591.00 (205)	[240, 1214] [258, 1344]	$U = 672.500$ $p = .918$	$r = .012$
# of tasks/questions completed per minute	(0) $n = 51$ (1) $n = 26$	3.71 (1.48) 3.58 (1.10)	[1.98, 5.63] [1.79, 5.14]	$U = 605.500$ $p = .536$	$r = -.071$
# of 10+ minute study days	(0) $n = 51$ (1) $n = 26$	12.00 (8) 11.00 (8)	[11.44, 14.60] [9.50, 15.20]	$U = 586.500$ $p = .409$	$r = -.094$
# days where all awaiting review tasks were completed	(0) $n = 51$ (1) $n = 26$	9.00 (7.00) 9.00 (9.00)	[2.00, 20.00] [4.00, 34.00]	$U = 766.000$ $p = .266$	$r = .127$

Note: (0) Non-gamified Group, (1) Gamified Group; (Statistical evidence* p value) $^{**} = (0.01 \leq P < 0.05)$ Moderate evidence; (Effect size* z-derived r) $\bullet = .100 \sim .300$ Small effect size.

number of tasks by utilizing the software on more days. Although the number of active study days is not shown to be not statistically significant under the non-parametric Mann-Whitney U test $p = .114$, it was a major factor which contributed to “tasks per active study day” being statistically significant.

Table 8 shows that both the “non-gamified” group ($n = 37$) and the “gamified” group ($n = 15$) significantly increased their test score throughout the twelve weeks. Although both groups recorded a very strong effect, the gamified group’s effect was slightly stronger $r(35) = .618$ than the one for the non-gamified group $r(13) = .558$.

Table 9 shows the pre/posttest results of the participants who completed both tests ($n = 52$) which assessed meaning, form, and use knowledge from a careful selection of 19 out of the final 463 ranked word frequencies of the NAWL. Statistical significance and area of effect calculations were calculated using the score/point differences between pre/posttest scores between the gamified group ($n = 15$) and non-gamified group ($n = 37$). The overall results showed the gamified group increased their scores significantly more than the non-gamified group; Mann-Whitney $U = 384.500$ $p = .03$, $r(50) = .300$ (see Figure 6). Moreover, an ANCOVA test for posttest results which controlled for pretest scores also produced similar results, $F = 4.321$ $p = .043$. However, as noted earlier, it would be better to not retain parametric test results because of data normality and distribution issues in many areas of this study including this one.

Table 8 Pre/Posttest Total Gains (within group)

Group#	Total Pretest Score/57 Median (IQR) [Min, Max]	Total Posttest Score/57 Median (IQR) [Min, Max]	Mann-Whitney Test	Effect size z-derived r
(0) $n = 37$	6.00 (8) [0, 16]	13.00 (8) [2, 44]	$U = 1128.000$ $p = .000^{****}$	$r = .558^{***}$
(1) $n = 15$	17.00 (10) [8, 30]	29.50 (9) [13, 42]	$U = 194.000$ $p = .000^{****}$	$r = .618^{***}$

Note: (0) Non-gamified Group, (1) Gamified Group; (Statistical evidence* p value) $^{****} = P < 0.001$ Strong evidence; (Effect size▪ z-derived r) $^{***} = .500 \sim .800$ Strong effect size.

Table 9 Gamification Level Comparison and Pre/Posttest Score Results

Group #	Test	Meaning score/19 [%] Median (IQR)	Form score/19 [%] Median (IQR)	Use score/19 [%] Median (IQR)	Total score/57 [%] Median (IQR)
Group (0) $n = 37$	Pre	[15.8%] 3.00 (4)	[10.5%] 2.00 (2)	[5.3%] 1.00 (2)	[10.5%] 6.00 (8)
	Post	[31.6%] 6.00 (4)	[26.3%] 5.00 (4)	[13.2%] 2.50 (3)	[22.8%] 13.00 (8)
	Diff.	[+15.8%] +3.00 (+0)	[+15.8%] +3.00 (+2)	[+7.9%] +1.50 (+1)	[+12.3%] +7.00 (+0)
Group (1) $n = 15$	Pre	[31.6%] 6.00 (4)	[31.6%] 6.00 (6)	[21.1%] 4.00 (4)	[29.8%] 17.00 (10)
	Post	[55.3%] 10.50 (3)	[52.6%] 10.00 (4)	[42.1%] 8.00 (4)	[51.8%] 29.50 (9)
	Diff.	[+23.7%] +4.50 (−1)	[+21%] +4.00 (−2)	[+21%] +4.00 (+0)	[+22%] +12.50 (−1)
Mann-Whitney		$U = 376.500$	$U = 328.500$	$U = 416.500$	$U = 384.500$
p value		$p = .045^{**}$	$p = .299$	$p = .005^{***}$	$p = .030^{**}$
z-derived r		$r = .278^{*}$	$r = .144^{*}$	$r = .393^{**}$	$r = .300^{*}$

Note: (0) Non-gamified Group, (1) Gamified Group [%] = Median%, Diff.= score difference; (Statistical evidence p value) $^{**} = (0.01 \leq p < 0.05)$ Moderate evidence; $^{***} = (0.001 \leq p < 0.01)$ Strong evidence; (Effect size z-derived r) $^{*} = .100 \sim .300$ Small effect size; $^{**} = .300 \sim .500$ Medium effect size.

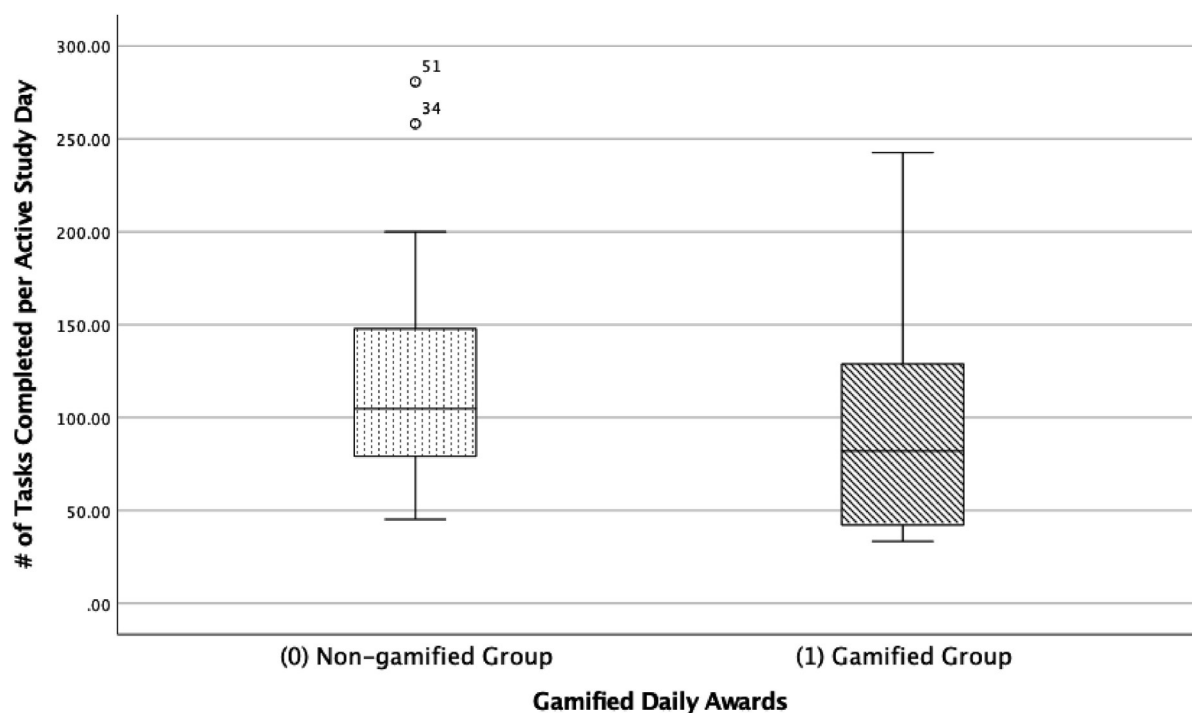


Figure 5 *Number of Tasks Completed Per Active Study Day Box Plot.*

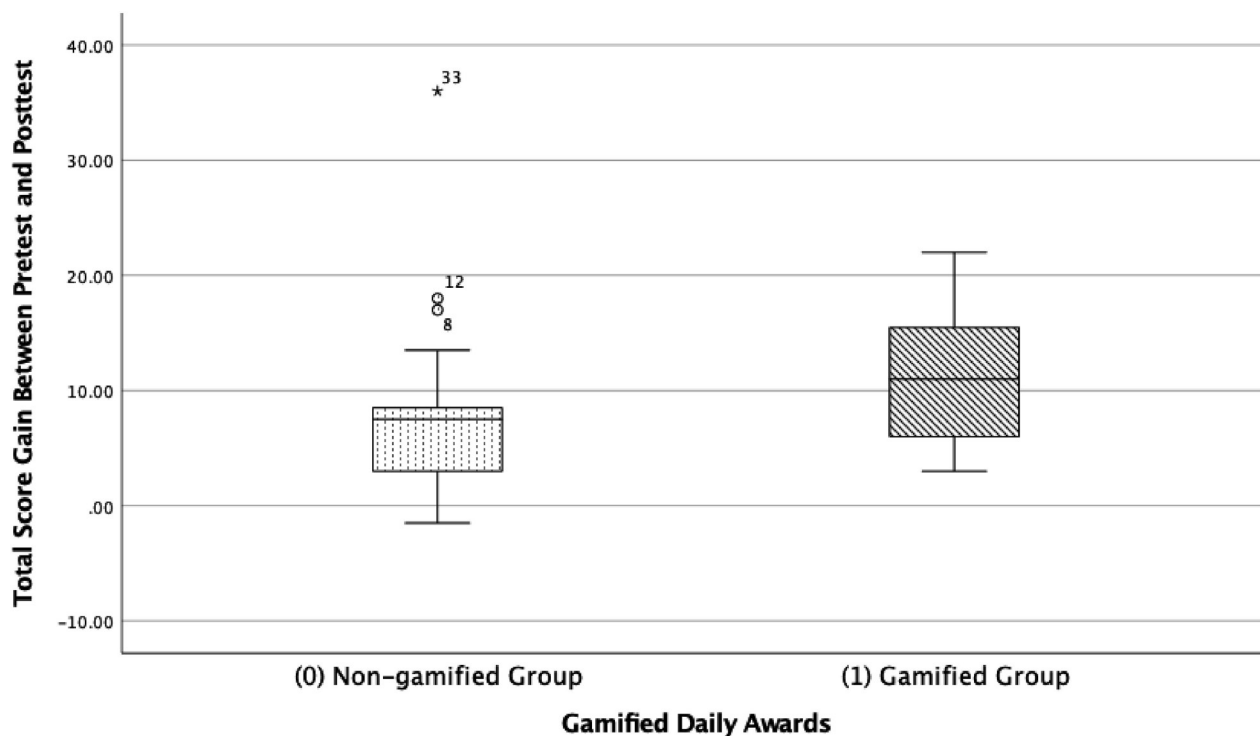


Figure 6 *Total Score Gain Between Pretest and Posttest Box Plot.*

Finally, NAWL vocabulary gain estimations revealed that the “non-gamified” group had roughly gained ~57 words, and the “gamified” group had gained ~102 words on average at the moment of the surprise/unannounced immediate posttest.

Discussion

RQ1: Do gamified daily awards have an effect on vocabulary software satisfaction?

As the data analysis shows, the group which had access to gamified daily awards outscored the group which did not in terms of satisfaction in all metrics: ease of use, interest, and usefulness (see Table 5). Moreover, the “gamified” group participants who had prior vocabulary learning software experience were more satisfied with the author’s software than with software they had used previously in contrast to the “non-gamified” group participants who were not (see Table 6). However, none of these results were statistically significant with the closest being software interest $p = .191$, $r(58) = .169$. Although this study did not reveal any area of statistical significance in terms of satisfaction, it would not be surprising if future studies with larger sample sizes and/or additional/better gamification elements disagreed with the results of this study.

RQ2: Do gamified daily awards have an effect on digital flashcard study habits?

The results showed that participants whose vocabulary learning software included gamified daily awards on average spread out their study efforts significantly more throughout the twelve-week study period in contrast to participants whose software did not have these daily gamified elements. The gamified group on average studied a lower number of tasks per active study day yet completed a similar number of tasks overall which points to the fact they used the software on more days (see Table 7). It could be hypothesized that such differences could have been even more pronounced if the study had not recommended a weekly study goal for all participants to follow. As far as total study time is concerned, the “non-gamified” group had a median of 580.00 minutes and the “gamified” group had a median of 591.00 minutes, which was not significant $U = 672.500$, $p = .918$. This slight yet non-significant difference could perhaps be attributed to the fact that each daily gamified award earned produced a pop-up animation that added a minimum of five seconds before the gamified group participant could click through them. Finally, it could be said that the gamified group’s study habits were more in line with the principles of spaced learning which encourage shorter but more numerous intervals of spaced study.

RQ3: Do gamified daily awards have an effect on vocabulary learning outcomes?

The results showed that participants whose vocabulary learning software included gamified daily awards had significantly better learning outcomes (i.e., most significant posttest score gain and estimated vocabulary increase; see Tables 8 & 9) than those whose software did not. Differences in terms of learning outcomes cannot be attributed to overall study time as there was no significant difference between the groups; Mann-Whitney $U = 672.500$, $p = .918$. Rather, the observed positive learning outcomes in regard to this study were in all probability due to the fact that the gamified elements were strongly informed by the principles of spaced learning that encourage numerous but shorter intervals of spaced study, which has been assessed as being more conducive to a higher efficiency of study in numerous studies (Kang, Lindsey, Mozer, & Pashler, 2014; Nakata, 2015; Pyc & Rawson, 2007). Moreover, the results of this study also echo the findings of Calvo-Ferrer (2017) and Fithriani (2021) regarding the advantage of gamified treatment group condition over the control group condition (i.e., non-gamified or traditional learning groups) in terms of vocabulary learning outcomes.

Conclusion

The promise of gamification in education has been identified by a number of researchers and pedagogic material developers as a way to significantly increase learners’ satisfaction and/or overall engagement. However, the results of this study questions if these are in fact the two most important

potential contributions. Indeed, perhaps the most important promise of gamification may actually be in improving learning efficiency (i.e., a higher ratio/number of items remembered per time invested) as observed in this study. These results likely were not due to random gamification features being included in the research project but most likely the result of the careful inclusion of gamification elements that were inspired by and promoted the principles of spaced learning. The gamified daily awards and bonuses included in the treatment group guided its participants to engage in study habits that were more conducive to learning (i.e., they were less prone to engage in cramming as they spread out their study efforts throughout the study period). These results present a compelling argument for educators and software developers to include daily gamified awards as a component of their educational software and/or pedagogical practices.

Limitations and Future Directions

Although this study was able to provide some insight into the specific research area of gamified daily awards, it should be noted that it also has its share of limitations that should be addressed in future studies. Firstly, in an effort to uphold a higher degree of testing scrutiny, it would be best to implement a multiple test rater approach or use a tried and tested automated testing system such as vocabularytest.org which agreed at a 98% rate with human marked meaning-recall responses (McLean et al., 2021). Moreover, the limited number of participants and the unevenness of their distribution in number and also L2 proficiency level within both study groups contributed to data issues (e.g., potentially less accurate p values and normality of distribution issues in the data). These limitations were controlled for by using non-parametric means of analysis such as the Mann-Whitney U test for calculating p values, z -derived r calculations for calculating effect sizes, and finally pre/posttest gain scores were preferred to pre/posttest results when calculating the effect of the software to control for varying participant proficiency levels. Although these various measures resolved or attenuated the effect of these issues, it would be best to avoid such problems in future studies from the onset by having larger and more evenly distributed participant sample sizes (i.e., in terms of both number and proficiency level) when conducting comparative studies.

Acknowledgments & Ethical Statement

The author would like to thank all those who contributed their advice and time to this paper, including the Japan Society for the Promotion of Science KAKENHI for their financial support which supported ISRS development, Grant-in-Aid for Scientific Research Grant Number JP19K00899 (<https://kaken.nii.ac.jp/en/grant/KAKENHI-PROJECT-19K00899/>).

The author conducted the research abiding by his institution's ethical regulations. Informed consent was obtained from all individual participants involved in the study. The author declares no conflicts of interest.

About the Author

Louis Lafleur is a lecturer at Kwansei Gakuin University conducting research related to the fields of second language vocabulary acquisition, cognitive psychology, computer-assisted language learning, and game-informed language learning (gamification).

References

- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>

- Browne, C., Culligan, B., & Phillips, J. (2013). *New Academic Word List (NAWL)*. Licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. New General Service List Project. Retrieved from <https://www.newgeneralservicelist.com/new-general-service-list-1>
- Calvo-Ferrer, J. R. (2017). Educational games as stand-alone learning tools and their motivational effect on L2 vocabulary acquisition and perceived learning gains. *British Journal of Educational Technology*, 48(2), 264–278. <https://doi.org/10.1111/bjet.12387>
- Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, 35(1), 3–25. <https://doi.org/10.1177/0265532216676851>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Dempsey, J. V., Barbara, L. A., Lucassen, L. L., Haynes, and Maryann, C. S. (1996, April). *Instructional applications of computer games*. Paper presented to the American Educational Research Association, New York, 8–12.
- Ebbinghaus, H. (1964). Über das gedächtnis: Untersuchungen zur experimentellen psychologie [Memory: A contribution to experimental psychology]. Smith (Original work published 1885).
- Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, 61(2), 367–413. <https://doi.org/10.1111/j.1467-9922.2010.00613.x>
- Fithriani, R. (2021). The utilization of mobile-assisted gamification for vocabulary learning: Its efficacy and perceived benefits. *Computer assisted Language Learning Electronic Journal (CALL-EJ)*, 22(3), 146–163. Retrieved from: <https://callej.org/index.php/journal/article/view/357/287>
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). “Effect size estimates: Current use, calculations, and interpretation”: Correction to Fritz et al. (2011). *Journal of Experimental Psychology: General*, 141(1), 2–18. <https://doi.org/10.1037/a0026092>
- George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. 11.0 update (4th ed.). Allyn & Bacon.
- Hughes, L. S. (2023a). Effects of dynamically computer-mediated communication and ‘pseudo’ communication on L2 Learning. *Computer Assisted Language Learning*, 24(1), 105–136. Retrieved from: <https://old.callej.org/journal/24-1/Hughes2023.pdf>
- Hughes, L. S. (2023b). The relationship between interaction moves during text-based SCMC and L2 vocabulary learning efficiency. *Technology in Language Teaching & Learning*, 5(1), 1–22. <https://doi.org/10.29140/tltl.v5n1.1045>
- Kanazawa, Y., & Lafleur, L. (2023). ENAWL: Enriching the New Academic Word List with emotional valence, familiarity, and knowledgeability. *Kokusaigaku Kenkyu–Journal of International Studies*, 12(1), 141–151. Retrieved June 7, 2023 from <http://hdl.handle.net/10236/00030725>
- Kang, S. H., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review*, 21(6), 1544–1550. <https://doi.org/10.3758/s13423-014-0636-z>
- Kapp, K. M. (2017). Gamification designs for instruction. In C. M. Reigeluth, B. J. Beatty, & R. D. Myers (Eds.), *Instructional-design theories and models, volume IV: The learner-centered paradigm of education* (pp. 351–384). Routledge. <https://doi.org/10.3758/s13423-014-0636-z>
- Lafleur, L. (2015). *The conceptualization of balanced and multifaceted vocabulary learning systems* (M. A.). Okayama University. Available on researchgate.net. <https://doi.org/10.13140/RG.2.2.22161.84327/1>
- Lafleur, L. (2020). The indirect spaced repetition concept. *Vocabulary Learning and Instruction*, 9(2), 9–16. <https://doi.org/10.7820/vli.v09.2.lafleur>
- Leitner, S. (1972). So lernt man lernen: Der weg zum erfolg [How to learn to learn: The road to success]. Freiburg im Breisgau, Baden-Württemberg: Verlag Herder.

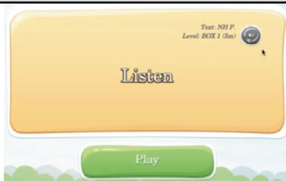
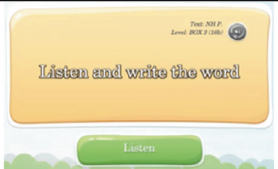
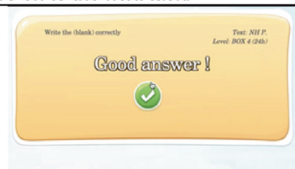
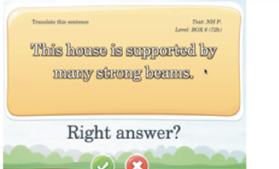
- McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, 37(3), 389–411. <https://doi.org/10.1177/0265532219898380>
- McLean, S., Raine, P., Pinchbeck, G., Huston, L., Kim, Y., Nishiyama, S., & Ueno, S. (2021). The internal consistency and accuracy of automatically scored written receptive meaning-recall data: a preliminary study. *Vocabulary Learning and Instruction*, 10(2), 64–81. <https://doi.org/10.7820/vli.v10.2.mclean>
- Mekler, E. D., Brühlmann, F., Tuch, A. N., & Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, 71, 525–534. <https://doi.org/10.1016/j.chb.2015.08.048>
- Nakata, T. (2008). English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning. *ReCALL*, 20(1), 3–20. <https://doi.org/10.1017/S0958344008000219>
- Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, 37(4), 677–711. <https://doi.org/10.1017/S0272263114000825>
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. <https://doi.org/10.1017/S0008413100018260>
- Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, 35(8), 1917–1927. <https://doi.org/10.3758/BF03192925>
- Reinhardt, J., & Sykes, J. M. (2012). Conceptualizing digital game-mediated L2 learning and pedagogy: Game-enhanced and game-based research and practice. In *Digital games in language learning and teaching* (pp. 32–49). Palgrave Macmillan. https://doi.org/10.1057/9781137005267_3
- Reinhardt, J., & Sykes, J. M. (2014). Guest editor commentary. *Language Learning & Technology*, 18(2), 2–8.
- Reinhardt, J. (2019). *Gameful second and foreign language teaching and learning: Theory, research, and practice*. New Language Learning and Teaching Environments. Palgrave Macmillan, Cham. <https://doi.org/10.1007/978-3-030-04729-0>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78. <https://doi.org/10.1037/0003-066X.55.1.68>
- Sundqvist, P., & Sylvén, L. K. (2016). *Extramural English in Teaching and Learning*. Palgrave Macmillan. <https://doi.org/10.1057/978-1-137-46048-6>
- Wang, Q. (2019). Classroom intervention for integration simulation games into language classrooms: An exploratory study with the SIMs 4. *CALL-EJ*, 20(2), 101–128. Retrieved from: <https://callegj.org/index.php/journal/article/view/275/206>

Appendix 1 Interleaved Spaced Repetition Software ISRS Tasks (adapted from Lafleur, 2020)

Tier Q# level	Flow L1 = native language L2 = second language	Task Type	Task/Interval Route: ↓ when answered successfully ↺ or ← when answered unsuccessfully (cooldown time/next review; m = minute, h = hour, d = day)		
Meaning Q#1 word or phrase	L2 audio to L1 recall	Recall Check	Session 1 (Q#1) ↺ (start/↺=5m) ↓	Session 7 (Q#1) ← (6d) ↓	Session 13 (Q#1) ← (162d) ↓
Meaning Q#2 word or phrase	L1 word/phrase to L2 recall	Recall Check	Session 2 (Q#2) ↺ (8h) ↓	Session 8 (Q#2) ← (9d) ↓	Session 14 (Q#2) ← (243d) ↓
Form Q#3 word or phrase	L2 audio to L2 word/phrase	Spelling	Session 3 (Q#3) ↺ (16h) ↓	Session 9 (Q#3) ← (18d) ↓	Session 15 (Q#3) ← (486d) ↓
Form Q#4 sentence	L2 (blank) to L2 sentence	Fill the blank	Session 4 (Q#4) ↺ (1d) ↓	Session 10 (Q#4) ← (27d) ↓	Session 16 (Q#4) ← (729d) ↓
Use Q#5 sentence	L2 sentence to L1 sentence	Writing	Session 5 (Q#5) ↺ (2d) ↓	Session 11 (Q#5) ← (54d) ↓	Session 17 (Q#5) ← (1458d) ↓
Use Q#6 sentence	L1 sentence to L2 sentence	Writing	Session 6 (Q#6) ↺ (3d) ↓	Session 12 (Q#6) ← (81d) ↓	Session 18 (Q#6) ← (2187d) end
(Optional)* Q#7 Text	L2 Listening and L2 Reading	(Voiced) Reading	↵ back to top ↑ ↵ back to top ↑ *optional, completing a set of words could trigger Q#7		

Note: Only one task/question type (Q#) is shown/asked with each study/review session; after a successful answer follow the ↓ arrow or ↺ / ← arrow after an unsuccessful answer; see Appendix 2 for question/task details.

Appendix 2 Interleaved Spaced Repetition Tasks

Q# Type	Detailed Task Flow		
Q#1 Meaning Word focus L2 audio to L1 recall	 1. "Listen" is displayed. 2. The user must push on the (play) button.	 3. "Think about the meaning of the word" is displayed. 4. The sentence/word audio are played. 5. The user thinks and when ready clicks on (check answer).	 6. Recommended answers are displayed. 7. The user self-assesses the validity of their answer/recall. 8. The user chooses right or wrong (honor system).
Q#2 Meaning Word focus L1 word to L2 recall	 1. "Think of the meaning (translate)" is displayed. 2. The L1 word(s)/synonyms are displayed. 3. The user thinks about a valid corresponding L2 word, and clicks on (see answers)	 4. Both recommended and viable answers are displayed. 5. The user self-assesses the validity of their answer/recall 6. The user validates or refutes their answer (honor system)	
Q#3 Form Word focus L2 audio to L2 word	 1. "Listen and Write the word" is displayed. 2. The user must push on the (listen) button.	 3. "Write the word" is displayed. 4. The sentence/word audio are played. 5. The user thinks and writes/spells the word and clicks (submit).	 6. The software automatically checks the user's answer. 7. "Good or wrong answer" is displayed. 8. The user must click on the flashcard to move on to the next task.
Q#4 Form Sentence focus L2 (blank) to L2 sentence	 1. "Write the blank (correctly)" is displayed. 2. The target sentence with a (blank) and L1 hints are shown.	 3. The user thinks about the answer and writes the missing word. 4. The user then clicks on (submit).	 5. The correct answer is shown. 6. The system automatically compares their answer with viable answers. 7. A click on screen is necessary to move on to the next task
Q#5 Use Sentence focus L2 sentence to L1 sentence	 1. "Translate this sentence" is displayed. 2. The target L2 sentence is displayed. 3. The user thinks about and writes the sentence in their L1 and clicks (submit).	 4. A recommended answer is displayed. 5. The user validates or refutes their answer (honor system).	
Use Q#6 Sentence focus L1 sentence to L2 sentence	 1. "Translate this sentence" is displayed. 2. The target L1 sentence is displayed. 3. The user thinks about and writes the sentence in their L2 and clicks (submit).	 4. A recommended answer is displayed. 5. The user validates or refutes their answer (honor system).	

Appendix 3 Pre/Posttest (with filled-in suggested/common answers for questions 1~19) New Academic Word List Test (token items for 501~963 range)

(Meaning)	(Form)	(Use)
例[example]. 教科書	textbook	その歴史の教科書には多くの間違いがあった。
1. 歴史的に; 歴史に関して	historically	私が発見した鏡は歴史的に重要であることが分かった。
2. 板挟み状態; 二律背反	dilemma	板挟み状態でストレスが多くたまった。
3. 洗練させる; 凝る; 複雑	sophisticate	残念ながら、彼女はあまり洗練されていない。
4. 影響されやすい; 感染しやすい	susceptible	彼女はウイルスに感染しやすい。
5. 安売り; 契約; 安い買い物	bargain	これは損な買い物[貧乏くじ]だ。
6. 持続可能な; 維持できる	sustainable	持続可能な開発は私たちの未来を救うだろう。
7. 矛盾する; 反対の	contradictory	この良い先生は自分が言うことに反することは行なわなかった。
8. 弾力のある; 融通の利く	elastic	この輪ゴムは十分に弾力があってしっかりしている。
9. 罰する; 乱暴に扱う	punish	彼は罰せられるべき人びとを的確に特定した。
10. 加工業者; 処理装置, プロセッサー	processor	教授が新しい処理装置の開発に成功した。
11. 余り; 剰余金; 黒字(金額)	surplus	私たちの会社は黒字[余剰]から赤字[不足]に転じた。
12. 毛管の; 毛状の; 毛細血管	capillary	喫煙が彼女の毛細血管を傷つけた。
13. 講義者; 講師	lecturer	私たちはとても人気のある講演者[講師]を招待した。
14. 社会化する; 社交的にする	socialize	多くの社長と社交したため、彼は良い仕事を見つけることができた。
15. 信用できること; 信用性	credibility	彼女の話は信憑性が高い。
16. モグラ; スパイ; ほくろ; 防波堤	mole	モグラ[スパイ]によって秘密が漏洩した。
17. 巧妙な; ずるい; 賢い; 難しい	tricky	そのドアは一筋縄では開かないので彼女はいらいらした。
18. こする; 摩擦する; する; 磨く	rub	彼が目をこすったときに危険なウイルスが体に入った。
19. 噴霧器; アトマイザー; 煙霧質	aerosol	有害なエアロゾル[空気中の煙霧質微粒子]は完全に除去された。