

Abstract

LAFLEUR LOUIS

The Effects of Interleaved Spaced Repetition and Gamification
on Digital Flashcard Vocabulary Learning in Foreign Language Education

This thesis consists of three studies that are related to Second Language Vocabulary Acquisition (SLVA) which has become its own important and popular research field within Second Language Acquisition (SLA). These studies entail important pedagogical implications in the area of digital flashcard vocabulary learning in foreign language education. The first, a pilot study, which collected survey data from ESL/EFL teachers in Japan, lays the foundation for the two main studies in terms of vocabulary teaching/learning considerations. The second study investigates the effects of Interleaved Spaced Repetition Software (ISRS) developed by the author for the purposes of vocabulary acquisition. The third and final study compares two versions of the author's software, one which included daily gamified awards, and the other which did not. All three studies identified and addressed gaps in the research literature, implemented a mixed-methods approach in collecting and analyzing both quantitative and qualitative data from their participants, and most importantly, report evidence that holds important pedagogical implications in the area of digital flashcard vocabulary learning in foreign language education.

The first study addresses the lack of research related to teaching practices and perspectives regarding word count units (WCU; e.g., lemma, word family). This study, conducted through an extensive online questionnaire, collected both quantitative and qualitative data from 30 ESL/EFL teachers in Japan. The questionnaire surveyed their vocabulary teaching experience and views regarding five categories of WCUs: All Forms and Meanings (AFM), plemma, lemma, flemma, and Word Family (WF). The online survey revealed that most participants had limited experience and knowledge regarding WCUs and

their impact. Another important finding is that although the WCU category participants were most familiar with is the WF (i.e., root/basic word lists), the WCU they had the highest pedagogical regard for is the lemma which considers parts of speech (POS) in vocabulary instruction and learning, followed closely by the plemma which additionally considers the various meanings words can take on, and also irregular inflectional patterns (e.g., irregular verb past tenses and past participles), and to a lesser extent the flemma which does not consider POS nor the various meanings and irregular inflectional patterns of words in vocabulary instruction.

The second study explored the learning of various aspects of word knowledge. Nation (2001) identified nine aspects of word knowledge which can be subsequently broken into receptive and productive areas of mastery for a more balanced approach in teaching and learning vocabulary. Interleaved Spaced Repetition Software (ISRS; i.e., digital flashcard study software) combines the principles of task interleaving (i.e., the reoccurring practice or study of multiple skills or concepts; e.g., the various aspects of word knowledge) and spaced repetition (i.e., interval-based study). This study which utilized a pretest-treatment-posttest data collection process to assess vocabulary learning outcomes enrolled a total of 74 Japanese university EFL learners over two academic semesters. During the treatment phase, the participants studied the New Academic Word List's (NAWL; Browne et al., 2013b) word-items supplemented with contextualized sentences, word/sentence audio and L1 translations (eNAWL; Kanazawa & Lafleur, 2023) with ISRS developed by the main author. Significant findings included the uneven pretest knowledge scores across the various tiers of knowledge: "Meaning" (24.34%), "Form" (20.53%) and "Use" (13.16%), and the relatively even posttest score gains as encouraged by ISRS's task interleaving algorithm: "Meaning" (+16.71%), "Form" (+15.39%), and "Use" (+13.16%). Finally, the treatment group significantly outperformed the control group in total score gains ($p = .002$, $r = .360$). These results present a reasonable argument regarding the viability of interleaved spaced repetition systems within the

area of vocabulary teaching and learning. Although this study was able to provide some insight into the research area of interleaved spaced repetition for vocabulary learning, it should be noted that it also has its share of limitations that should be addressed in future studies. Although some precautions were taken to uphold a high degree of testing scrutiny (e.g., the main author was blind to which group tests were associated while grading), it would be best to implement a multiple-test-rater approach in future studies. Unfortunately, in the case of this study, although efforts were made to find raters, none accepted the author's proposition. Second, the limited number of participants and the unevenness of their L2 proficiency level within both study groups contributed to data issues (e.g., normality of distribution issues in the data). These limitations were controlled for by using non-parametric means of analysis. It would be best to avoid such issues 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). Finally, many avenues of research are still left unexplored, for example, future studies could compare ISRS to non-interleaved/traditional spaced repetition software, compare the efficiency of various types of expanding spaced algorithms, and explore the study/learning of other study content.

The third study investigates the effects of daily gamified awards in second language acquisition research. This novel 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) 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. 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 vocableveltest.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-/post-test gain scores were preferred to pre-/post-test results when calculating the effect of the software in an effort 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.

This research yields evidence regarding the pedagogical benefits of Interleaved Spaced Repetition Software (ISRS) and the positive effects of daily gamified awards on digital flashcard vocabulary learning in foreign language education. Moreover, it also offers new avenues and perspectives to guide future research.

(1253 words)