ISSN:2828-6235
Impact factor:13.77

International Journal of Literature and Language Studies(IJLLS)

GAMIFIED DIGITAL TOOLS AND VOCABULARY RETENTION IN EFL CLASSROOMS: EFFECTS ON LEARNING OUTCOMES AND MOTIVATION AMONG FIRST-YEAR UNIVERSITY STUDENTS IN UZBEKISTAN

Olloyorova Jasmina To'lqin qizi

Doi: https://doi.org/10.5281/zenodo.23570874

Abstract: Vocabulary knowledge that is not retained over time contributes little to communicative competence, yet retention remains a persistent challenge in English as a foreign language (EFL) classrooms. This quasi-experimental, mixed-methods study examined whether gamified digital tools (Kahoot! and Quizlet) improve vocabulary retention and learning motivation among first-year students at a state university in Uzbekistan. Sixty students aged 18–20 were assigned by intact class to an experimental group (n = 30) or a control group (n = 30). Over six weeks, both groups studied the same 60 target words; the experimental group practised through gamified platforms, whereas the control group used non-gamified paper-based activities. Learning was measured by a pre-test, an immediate post-test, and a delayed post-test administered four weeks later, and motivation by a 5-point Likert questionnaire. Analysis combined ANCOVA, independent-samples t-tests, and effect sizes. In the illustrative model data reported here, the experimental group outperformed the control group on both post-tests and reported higher motivation. The article discusses the findings in relation to depth of processing, retrieval practice, and self-determination theory, and outlines pedagogical implications and limitations.

Keywords: gamification; vocabulary retention; EFL; Kahoot!; Quizlet; learner motivation

INTRODUCTION

Vocabulary knowledge is widely regarded as a foundation of second language proficiency, because learners who lack a sufficient lexical base struggle to comprehend input and to express their own meanings (Nation, 2001; Schmitt, 2008). Acquiring a word is not a single event but an incremental process in which form, meaning, and use are consolidated through repeated, meaningful encounters. For many first-year university students in Uzbekistan, English is a foreign language with limited exposure beyond the classroom, and in the authors' experience instruction often still relies on word lists, translation, and end-of-unit tests. Such practice can produce short-term recognition of words that is rapidly forgotten, which undermines later reading, writing, and speaking tasks.

At the same time, students entering university are habitual users of smartphones and games, and teachers increasingly use digital platforms such as Kahoot! and Quizlet to review vocabulary. These tools add points, leaderboards, timers, and competition to familiar practice activities, and their popularity suggests that they may raise engagement. Whether this engagement translates into durable lexical learning, rather than a short-lived novelty effect, is an open empirical question, especially in Central Asian higher-education contexts where controlled evidence is scarce. The present study addresses this gap by comparing gamified and non-gamified vocabulary practice, measuring both immediate and delayed outcomes, and examining learner motivation alongside test performance.

Deterding et al. (2011) define gamification as the use of game design elements in non-game contexts. In language education, such elements (points, badges, levels, leaderboards, and

immediate feedback) are layered onto otherwise ordinary practice. Three theoretical perspectives help explain why this might affect vocabulary learning. First, the levels-of-processing framework holds that memory durability depends on the depth at which information is processed (Craik & Lockhart, 1972). In vocabulary research this idea is elaborated in the involvement load hypothesis, which proposes that tasks combining need, search, and evaluation lead to better retention than tasks that demand less cognitive involvement (Hulstijn & Laufer, 2001). Gamified quizzes may increase involvement by making the need to retrieve a word salient and by requiring rapid evaluation of alternatives.

Second, retrieval and spacing research indicates that being tested on material, rather than simply re-reading it, improves long-term retention (Roediger & Karpicke, 2006), and that distributing practice over time benefits recall more than massing it (Cepeda et al., 2006). Platforms such as Quizlet and Kahoot! are built around low-stakes, repeated retrieval, which may explain part of their pedagogical value independently of their game features. Third, self-determination theory proposes that intrinsic motivation flourishes when the needs for autonomy, competence, and relatedness are satisfied (Ryan & Deci, 2000). Visible progress, immediate feedback, and team or class competition may support competence and relatedness, while the sustained effort that vocabulary learning requires is consistent with Dörnyei's (2001) view that motivation must be actively generated and maintained in the language classroom.

These perspectives converge with the practical account of vocabulary learning offered by Nation (2001), who emphasises that a word is learned through repeated encounters that involve its form, meaning, and use, and by Schmitt (2008), who argues that engagement with target items, and not merely exposure to them, is a key condition for instructed vocabulary learning. Gamified tools are therefore best understood not as a substitute for these conditions but as a means of increasing the quantity and quality of engagement within them.

Reviews of gamification research generally report positive but context-dependent effects. Hamari et al. (2014) concluded that gamification often produces the intended outcomes, but that results depend on the context and on the users involved. A later meta-analysis found small-to-moderate positive effects of gamified instruction on cognitive, motivational, and behavioural outcomes, with substantial variation across studies (Sailer & Homner, 2020). Within language learning, a systematic review by Dehghanzadeh et al. (2021) suggested that gamified applications tend to support learning and engagement, while noting methodological weaknesses in many primary studies. Research on Kahoot! specifically indicates favourable effects on performance and learner attitudes, although the literature is dominated by short interventions (Wang & Tahir, 2020).

Three limitations recur in this body of work. Many studies rely on immediate post-tests and therefore cannot distinguish genuine retention from short-term enthusiasm; control conditions frequently differ from experimental conditions in time on task or in the amount of exposure to target items; and relatively few studies examine learning outcomes and motivation together in the same sample. Little of the available evidence concerns first-year students in Uzbekistan. The present study responds by holding the target words, instructional time, and teacher constant across groups, by adding a delayed post-test, and by collecting motivational data in a mixed-methods design.

1.3 Research questions

The study addressed three research questions, each paired with a directional hypothesis.

RQ1. Does gamified digital vocabulary practice lead to higher immediate post-test scores than non-gamified practice, after controlling for pre-test scores? *H1:* The experimental group will outperform the control group on the immediate post-test.

RQ2. Does gamified practice lead to higher vocabulary retention four weeks after instruction? *H2:* The experimental group will outperform the control group on the delayed post-test.

RQ3. Do students who use gamified tools report higher learning motivation than students in the control group? *H3:* Motivation scores will be higher in the experimental group.

METHODS

The study used a mixed-methods quasi-experimental design with a pre-test, immediate post-test, delayed post-test, and non-equivalent control group (Creswell & Plano Clark, 2018). The quantitative strand compared achievement and motivation between groups; a brief qualitative strand, consisting of two open-ended questions, helped to interpret the motivation scores. Participants were 60 first-year undergraduates aged 18–20 enrolled in a compulsory EFL course at a state university in Uzbekistan. Because random assignment was not possible, two intact classes were allocated to the experimental group (n = 30) or the control group (n = 30) by a coin toss at class level. Both classes were at a comparable pre-intermediate level of English according to the department's placement results and were taught by the same instructor. A sensitivity analysis indicated that groups of this size provide 80% power to detect a standardised difference of about d = 0.74 (two-tailed, α = .05).

Sixty target words were selected from the second and third thousand-word frequency bands (Nation, 2001) and from the course textbook. Words that most students recognised in a screening list were excluded so that the items represented genuinely new learning. Ten words were taught each week, with nouns, verbs, and adjectives balanced across sets.

Vocabulary tests. The same 60 items were used in a pre-test, an immediate post-test (the final lesson of week 6), and a delayed post-test (four weeks later, with no announcement and no further instruction on the words). Each item was a context sentence with a gap, the first letter of the missing word provided; one point was awarded for a correctly spelled target word (maximum score = 60). Item order was re-randomised at each administration, and the scoring was completed by two raters who resolved disagreements by discussion. Testing recall of form in context was chosen because it requires productive retrieval, which is a more demanding criterion than recognition (Schmitt, 2008).

Motivation questionnaire. A 12-item questionnaire on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree) was adapted from motivational constructs described by Dörnyei (2001) and Ryan and Deci (2000). It contained three four-item subscales: interest and enjoyment, effort and persistence, and confidence. The instrument was administered in the students' first language at the end of week 6, and two open-ended questions asked what students liked and disliked about the vocabulary activities.

The study lasted ten weeks. The pre-test was given in week 0, the intervention ran for six weeks (two 80-minute lessons per week), the immediate post-test and questionnaire followed in week 6, and the delayed post-test took place in week 10. In both groups, the weekly set of ten words was introduced in the first lesson through the same short reading text and teacher explanation. Approximately 15 minutes per lesson, plus a short weekly homework task, were then devoted to practice, and this time was identical across groups.

The experimental group practised the words with gamified platforms. Teacher-created Quizlet sets supported flashcard study, Learn, and Match modes, and students accessed them on their smartphones or in the computer laboratory for homework. Each lesson ended with a Kahoot! quiz in which students answered timed multiple-choice and true/false items on the target words, earned points for speed and accuracy, and saw a leaderboard; teams were used in alternate weeks to encourage cooperation. The control group received the same words, definitions, and number of retrieval opportunities through paper-based matching exercises, gap-fill worksheets, oral questioning, and teacher-led review. Control activities were deliberately free of points, rankings, and competition, so that the main contrast was the presence of game elements and digital delivery.

Data were analysed with SPSS. Baseline equivalence was checked with an independent-samples *t*-test on pre-test scores. Group differences on the immediate and delayed post-tests were tested with one-way ANCOVA using the pre-test as covariate, after checking normality (Shapiro–Wilk), homogeneity of variance (Levene's test), and homogeneity of regression slopes. Independent-samples *t*-tests and 95% confidence intervals for raw mean differences were reported for descriptive clarity, together with Cohen's *d* and partial η^2. Following Cohen (1988), partial η^2 values of .01, .06, and .14 were interpreted as small, medium, and large. Questionnaire scores were compared with independent-samples *t*-tests and verified with Mann–Whitney *U* tests because Likert data are ordinal; internal consistency was estimated using Cronbach's α. Open-ended responses were coded inductively, and recurrent themes were counted. The significance level was set at .05.

Because each hypothesis addressed a different outcome and was tested once, no correction for multiple comparisons was applied to the three primary tests; the four subscale comparisons in the questionnaire were treated as exploratory. All tests were two-tailed, and effect sizes and confidence intervals were emphasised over significance values alone.

Participation was voluntary, and informed consent was obtained from all students after the aims and procedures were explained. Test and questionnaire data were anonymised with numerical codes and used for research purposes only. Students in the control group were offered access to the digital study sets after the delayed post-test so that they were not disadvantaged. [Insert institutional ethics approval or department permission reference.]

RESULTS

NOTE FOR AUTHOR: The figures in this section are illustrative model data created to show the reporting format. Replace them with the actual data from your own study before submission.

All 60 students completed all three tests, so no data were missing. Test reliability was high (Cronbach's α = .91 for the pre-test, .92 for the immediate post-test, and .93 for the delayed post-test). Shapiro–Wilk tests did not indicate departures from normality in either group, and Levene's test showed equal variances for the delayed post-test, $F(1, 58) = 0.00$, $p = .951$. The groups did not differ at baseline, $t(58) = -0.35$, $p = .730$, $d = -0.09$ (Table 1).

Both groups improved markedly from pre-test to post-test, but the experimental group's gains were larger. Table 1 shows that the experimental group scored 47.23 on average ($SD = 6.27$) on the immediate post-test, compared with 39.63 ($SD = 6.30$) for the control group. Four weeks later the experimental group still averaged 41.87 words ($SD = 6.69$), whereas the control group

averaged 31.47 (*SD* = 6.39). From immediate to delayed testing, the experimental group lost 5.37 points on average and the control group lost 8.17 points.

Table 1

Descriptive Statistics for Vocabulary Test Scores (Maximum = 60)

Test	Group	*n*	*M*	*SD*	Range
Pre-test	Experimental	30	21.43	5.11	12–31
	Control	30	21.90	5.30	11–32
Immediate post-test	Experimental	30	47.23	6.27	35–58
	Control	30	39.63	6.30	27–52
Delayed post-test	Experimental	30	41.87	6.69	29–54
	Control	30	31.47	6.39	19–44

Note. Scores represent the number of target words correctly produced. Illustrative model data.

Independent-samples *t*-tests (Table 2) showed no baseline difference but significant, large differences favouring the experimental group on the immediate post-test, $t(58) = 4.68$, $p < .001$, $d = 1.21$, and the delayed post-test, $t(58) = 6.16$, $p < .001$, $d = 1.59$. Paired comparisons confirmed that both groups improved from pre-test to immediate post-test, experimental $t(29) = 28.87$ and control $t(29) = 17.15$, both $p < .001$.

Table 2

Independent-Samples t-Tests Comparing the Experimental and Control Groups

Test	Mean difference	*t*(58)	*p*	Cohen's *d*	95% CI
Pre-test	−0.47	−0.35	.730	−0.09	[−3.16, 2.22]
Immediate post-test	7.60	4.68	< .001	1.21	[4.35, 10.85]
Delayed post-test	10.40	6.16	< .001	1.59	[7.02, 13.78]

Note. Mean difference = experimental minus control. CI = confidence interval for the mean difference. Illustrative model data.

The ANCOVA results (Table 3) were consistent. After adjustment for pre-test scores, group membership explained a significant proportion of variance in immediate post-test scores, $F(1, 57) = 35.93$, $p < .001$, partial $\eta^2 = .39$, and in delayed post-test scores, $F(1, 57) = 52.29$, $p < .001$, partial $\eta^2 = .48$, both large effects. Adjusted means were 47.40 versus 39.47 on the immediate post-test and 42.01 versus 31.32 on the delayed test. H1 and H2 were therefore supported. The groups did not differ significantly in the amount of forgetting between the two post-tests, $t(58) =$

−1.49, p = .142, d = 0.38, which indicates that the delayed advantage largely reflected a higher level of initial learning, with a non-significant tendency toward better retention in the experimental group.

Table 3

ANCOVA Results for Post-Test Scores With Pre-Test as Covariate

Outcome	$F(1, 57)$	p	Partial η^2	Adj. M (Exp.)	Adj. M (Con.)
Immediate post-test	35.93	< .001	.39	47.40	39.47
Delayed post-test	52.29	< .001	.48	42.01	31.32

Note. Adj. M = covariate-adjusted mean; Exp. = experimental group; Con. = control group. Illustrative model data.

The questionnaire showed good internal consistency (α = .88). The experimental group reported higher overall motivation (M = 4.18, SD = 0.51) than the control group (M = 3.52, SD = 0.60), $t(58) = 4.58$, $p < .001$, $d = 1.18$; a Mann–Whitney test gave the same conclusion, $U = 723$, $p < .001$. Differences were comparable across the three subscales (Table 4), supporting H3.

Table 4

Motivation Questionnaire Scores by Subscale (5-Point Likert Scale)

Subscale	Experimental M (SD)	Control M (SD)	$t(58)$	p	d
Interest and enjoyment	4.31 (0.55)	3.64 (0.66)	4.27	< .001	1.10
Effort and persistence	4.12 (0.58)	3.47 (0.63)	4.16	< .001	1.07
Confidence	4.11 (0.62)	3.45 (0.68)	3.93	< .001	1.01
Overall	4.18 (0.51)	3.52 (0.60)	4.58	< .001	1.18

Note. Each subscale contains four items. Illustrative model data.

In the open-ended responses, experimental-group students most often mentioned that quizzes were enjoyable and "not boring" and that immediate feedback showed them which words they still did not know; a minority reported anxiety about being ranked publicly or difficulty with unstable internet access. Control-group students more often described the practice as useful but repetitive. These comments are illustrative and are intended to show how the qualitative strand would be summarised.

DISCUSSION

The model results indicate that gamified digital practice was associated with higher vocabulary scores both immediately and four weeks after instruction, together with higher self-reported motivation. This pattern is consistent with the broader literature in which gamification yields small-to-moderate average benefits (Sailer & Homner, 2020) and with evidence that Kahoot!-type activities favour performance and attitudes (Wang & Tahir, 2020). The effects

reported here are larger than those typically found in meta-analyses, which is a reason for caution; it should be borne in mind that a single-site study with intact classes may overestimate effects, and that the figures in this article are illustrative.

Several mechanisms may explain the advantage. Because the control group received the same words, time, and retrieval opportunities, the contrast cannot be attributed only to testing or spacing (Cepeda et al., 2006; Roediger & Karpicke, 2006). Instead, the time pressure, competition, and instant feedback of the platforms may have increased the involvement load of each encounter by making retrieval and evaluation more demanding and more consequential for the learner (Hulstijn & Laufer, 2001), thereby promoting deeper processing (Craik & Lockhart, 1972). The motivational results complement this account: the higher interest, effort, and confidence scores are consistent with a self-determination perspective, in which feedback and visible progress support competence and class competition supports relatedness (Ryan & Deci, 2000). The finding that forgetting between the two post-tests did not differ significantly suggests that the principal benefit lay in stronger initial encoding rather than in slower decay, a nuance that studies relying on immediate tests alone would miss.

The Uzbek higher-education context may also be relevant. Large groups, limited contact hours, and a tradition of teacher-centred instruction can leave students with few opportunities to retrieve new words actively and to receive feedback on them. Platforms that let every student answer every item simultaneously, and that show results immediately, may therefore offer a degree of individual participation that is hard to achieve through oral questioning alone. This interpretation is plausible but untested here, and it should be examined directly through observation and interviews before it is generalised.

For EFL teachers in Uzbekistan and similar contexts, the findings suggest that low-cost platforms can be integrated into existing syllabi without replacing sound vocabulary pedagogy. First, gamified activities should be built on principled practice: repeated, spaced retrieval of a limited set of words in meaningful contexts (Nation, 2001). Second, teachers should combine game elements with opportunities for productive use, since quizzes that stress recognition alone may not develop the ability to use words in speech and writing (Schmitt, 2008). Third, teams, rotating roles, and optional anonymity on leaderboards can protect less confident students from the anxiety some learners associate with public ranking. Finally, institutions should consider infrastructure realities, including device access and internet reliability, and provide offline alternatives such as printable Quizlet sets.

These implications also concern teacher development and assessment. Teachers need training not only in operating the platforms but also in designing items that target the appropriate depth of knowledge, since poorly written distractors can teach incorrect forms. Departments may wish to share vetted word sets across classes, and teachers should align periodic assessment with the type of knowledge that the games train, so that students perceive the activities as relevant to course outcomes rather than as entertainment.

The study has several limitations. Assignment by intact class rather than by individual randomisation leaves open the possibility of unmeasured class-level differences, although the baseline equivalence and ANCOVA adjustment mitigate this concern. The sample was small, drawn from one university, and taught by one instructor, which limits generalisability and confounds the intervention with instructor enthusiasm. The experimental condition differed from the control condition in both medium (digital versus paper) and game design, so the unique contribution of game elements cannot be isolated. Only one delayed test was given, the target words were tested in a single format, and the motivation measure relied on self-report adapted for

this study rather than a fully validated scale. Finally, the novelty of the technology may have inflated enthusiasm over a six-week period.

Future studies should randomise at the student or section level, include several institutions and instructors, and add a third condition that uses the same platforms without game elements, so that digital delivery and gamification can be separated. Longer follow-up periods of three to six months, multiple test formats covering receptive and productive knowledge, and measures of vocabulary use in writing and speaking would show whether benefits are durable and transferable. Combining questionnaires with classroom observation, interviews, and platform analytics would also clarify which game elements sustain motivation and for whom.

CONCLUSION

This article presented a design, a reporting format, and illustrative results for a study of gamified vocabulary practice among first-year EFL students in Uzbekistan. Under the model data, Kahoot! and Quizlet activities were associated with higher immediate and delayed vocabulary scores and with higher motivation than equivalent non-gamified practice. If replicated with actual data, such findings would support the principled use of gamified tools as a supplement to well-designed vocabulary instruction, with attention to inclusion, access, and productive use of new words. The main contribution of the design lies in its control of time on task and its inclusion of a delayed test, which together allow clearer conclusions about retention than studies based on immediate scores alone.

REFERENCES

1. Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380.
2. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
3. Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684.
4. Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.
5. Dehghanzadeh, H., Fardanesh, H., Hatami, J., Talaee, E., & Noroozi, O. (2021). Using gamification to support learning English as a second language: A systematic review. *Computer Assisted Language Learning, 34*(7), 934–957.
6. Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference* (pp. 9–15). ACM.
7. Dörnyei, Z. (2001). *Motivational strategies in the language classroom*. Cambridge University Press.
8. Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? A literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences* (pp. 3025–3034). IEEE.
9. Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*(3), 539–558.
10. Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.
11. Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

12. 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.
13. Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review, 32*(1), 77–112.
14. Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329–363.
15. Wang, A. I., & Tahir, R. (2020). The effect of using Kahoot! for learning: A literature review. *Computers & Education, 149*, Article 103818.