

Zhorabekova A.¹, *Duisenova M.², Kurmasheva M.³, Yesnazar A.⁴

¹ M. Auezov South Kazakhstan University

^{2,3,4} O. Zhanibekov South Kazakhstan Pedagogical University

^{1, 2,3,4} Kazakhstan, Shymkent

¹ORCID ID: 0000-0003-2697-8077

²ORCID ID: 0000-0001-5684-7364

³ORCID ID: 0000-0002-4965-8490

⁴ORCID ID: 0000-0002-8658-0135

*duisenova.marzhan@okmpu.kz

COMPARING AN ARTIFICIAL INTELLIGENCE-BASED ADAPTIVE APPLICATION AND A GAMIFIED APPLICATION FOR MOBILE ENGLISH LEARNING AMONG UNIVERSITY STUDENTS IN KAZAKHSTAN

Annotation

Mobile applications have become a central tool for learning English as a foreign language. Two designs now dominate the market, gamification and artificial-intelligence-based adaptive learning, yet they are seldom compared directly, especially in Kazakhstan and among the future foreign-language teachers who will shape its national language education. This study compared a gamified application, Duolingo, with an artificial-intelligence-based adaptive application developed in Kazakhstan, Oki/Aleem, over a four-week intervention. A two-group quasi-experimental design was used. In total, 120 undergraduate students in a foreign-language teacher-training programme worked with one of the two applications, and 115 were matched across a pre-test and a post-test. Achievement was measured with parallel twenty-item forms administered before and after the intervention. A weekly form tracked how the applications were used, and three open-ended questions were coded thematically. Scores rose markedly in both groups over the month. Once pre-test scores were entered as a covariate, the analysis of covariance showed no advantage for either application. Weekly practice was sustained in both groups and increased toward the end of the intervention. The open-ended responses revealed a clear contrast. Duolingo was valued for its game-like motivation, whereas Oki/Aleem was valued for its adaptivity, feedback and local relevance. The two applications therefore produced similar learning gains through different mechanisms, which reframes the choice between them as a question of purpose rather than of effectiveness. The findings support building direct, comparative experience of gamified and adaptive applications into foreign-language teacher preparation in Kazakhstan.

Keywords: artificial intelligence, English as a foreign language, gamification, Duolingo, Oki/Aleem, mobile-assisted language learning, foreign-language teacher preparation.

Introduction. Mobile applications have moved to the centre of how people learn English as a foreign language. Smartphones are now widely available and allow learners to study at almost any time and place, and mobile-assisted language learning has grown from a marginal practice into a mainstream one, supported by a large and still-expanding body of research [1]. For many learners, a language application is no longer a supplement to the classroom but the primary space for everyday practice [2].

Two design philosophies have come to dominate this field. The first is gamification, which uses points, badges, streaks, levels and daily goals to keep learners returning to the application. Its best-known representative is the global application Duolingo. The second is artificial-intelligence-based adaptive learning, in which the application analyses a learner's performance and adjusts tasks, feedback and difficulty to the individual. This approach underpins a newer generation of artificial-intelligence tutors. Both promise stronger motivation and measurable language gains, but they pursue these outcomes through very different mechanisms, and it is not obvious that they bring learners to the same result.

The question is especially salient in Kazakhstan. Under the national digitalisation agenda, the digitalisation of higher education has become a policy priority, and the preparation of future foreign-language teachers increasingly assumes that graduates will be competent users and designers of such tools [3]. A locally developed adaptive application, Oki/Aleem, has begun to compete with Duolingo for the attention of Kazakhstani learners, which adds cultural and curricular relevance to the

comparison. Teachers and students thus face a practical choice between a gamified and an adaptive application, yet they have little direct evidence on which to base it.

The present study addresses this gap. For four weeks, undergraduate students in a foreign-language teacher-training programme worked with either Duolingo or Oki/Aleem. We tracked their achievement gains and their weekly use of the applications, and we asked the students what each tool was like to work with. By pairing an objective achievement test with learners' own accounts, the study speaks both to the measurable effects of the two designs and to the reasons learners value them.

Literature review. Reviews of the mobile-assisted language learning literature generally report positive effects on language learning, but they also caution against reading those effects too confidently. The largest meta-analysis to date is that of Burston and Giannakou [1, p. 160], who pooled experimental studies published between 1994 and 2019. The overall effect sizes are large, approximately 0.72 in between-group designs and 1.16 in within-group designs, yet the authors warn that these values should be interpreted with caution because of clear publication bias and very high heterogeneity [1, pp. 161-162]. The same synthesis shows that vocabulary is by far the most frequently targeted learning focus, accounting for roughly a third of all reported outcomes, while advanced learners are rarely studied [1, p. 159]. Longitudinal evidence indicates that sustained mobile practice can shift learners' attitudes and self-regulation in a positive direction over time [2, p. 8]. The field therefore appears effective but uneven, with short interventions, small samples and non-standardised tests recurring as weaknesses.

Gamification, understood as the use of game-design elements such as points and rewards in non-game settings, has been introduced into foreign-language teaching to address the motivational problem at the heart of self-directed study [4]. Meta-analytic evidence is broadly favourable. In a meta-analysis of South Korean research, Lee and Baek reported a medium effect of gamification on English proficiency, with a standardised mean difference of about 0.52 [5]. This average conceals a moderator effect. The gain was largest in primary school, whereas at university level it was small and not statistically significant [5, p. 11]. A further caveat recurs across the reviews. Technology-based study is self-directed, so keeping learners engaged is a problem in its own right, and the features intended to motivate can also distract [6].

Duolingo, the application used by one group in this study, is the most heavily researched gamified platform. One study assessed 245 learners who had used Duolingo as their only learning tool and had completed its beginner content at level A2 of the Common European Framework of Reference for Languages (CEFR), using a standardised proficiency test [7]. On average these learners reached Intermediate High in both reading and listening, one sublevel of the American Council on the Teaching of Foreign Languages (ACTFL) scale above the expected course outcome [7, p. 263]. A natural experiment with several hundred learners showed that the regularity of practice and engagement with the curriculum predicted gains more dependably than the number of minutes spent in the application [8]. Comparative studies indicate that Duolingo can match classroom instruction. Beginner learners of French improved to a similar degree whether they studied in class, through Duolingo alone or through both combined, the single exception being one pragmatic measure on which the combined condition outperformed the other two [9]. What this literature rarely does is compare Duolingo with an adaptive, artificial-intelligence-based alternative rather than with classroom instruction or with itself.

The adaptive-learning approach rests on a different premise, namely that learning is most effective when content, difficulty and feedback are tailored to the individual in real time. Experiments testing this premise are only now appearing. One pre-test/post-test control-group study found that learners taught with adaptive artificial-intelligence-based instruction handled difficult grammatical structures better than those taught conventionally, although the sample was small ($n = 60$) [10]. Achievement is not the whole story. Corrective and motivational feedback generated by artificial intelligence has been associated with higher engagement, enjoyment and motivation [11]. Systematic reviews nevertheless note that the evidence base is still young and uneven. Reviewing studies published between 2017 and 2023, one review concluded that the integration of artificial intelligence into English teaching remains at an early stage and that rigorous evidence on its effects is limited

[12]. Adaptivity and rich feedback are therefore plausible strengths of an application such as Oki/Aleem, but they have seldom been tested directly against a gamified competitor.

The Kazakhstani setting matters as well. The national digitalisation policy has produced a wave of research on technology in higher education. These studies report that digital interventions can strengthen students' professional and cognitive skills, but they also expose gaps in infrastructure and in teacher readiness [13]. In teacher preparation specifically, a mixed-methods study of Kazakhstani teacher-training programmes showed that digital learning environments improve future English teachers' language proficiency, pedagogical skill and technological fluency [3, p. 284], while problems of access and of integrating technology into the curriculum remain unresolved [3, p. 285]. Since the digital competence of future language teachers is now treated as a core professional requirement, this cohort is a particularly relevant group in which to study how different application designs work in practice.

Across these strands the same gap recurs. The literature compares applications with classroom instruction [9, p. 22], or reviews a single application type, gamified or adaptive, on its own. It rarely places a gamified application and an adaptive application alongside each other within the same learner population, under comparable conditions, with the same prescribed dose and outcome measures. Head-to-head comparisons of this kind are rare, and we found almost none from Central Asia or Kazakhstan. A locally built application such as Oki/Aleem is, for practical purposes, invisible in the international evidence base. Future foreign-language teachers, the cohort on which the region's long-term language education depends, are also under-represented. The present study was designed to address this gap.

The aim and objectives of the study. The aim of this study is to compare the effects of a gamified mobile application (Duolingo) and an artificial-intelligence-based adaptive mobile application (Oki/Aleem) on the English achievement, application use and perceptions of undergraduate learners of English in a Kazakhstani teacher-training programme over a four-week intervention. The specific objectives are: (1) to measure and compare English achievement gains in the two groups; (2) to describe how consistently each group used its assigned application; (3) to examine, through learners' own accounts, how they perceived the usefulness, difficulties and relative suitability of the two applications; and (4) to consider what these findings imply for the use of gamified and adaptive applications in English learning and in foreign-language teacher preparation in Kazakhstan.

These objectives were addressed through four research questions and four hypotheses (H). RQ1: after a four-week intervention, do learners in the two groups differ in their English achievement gains? (H1: both groups will improve significantly from pre-test to post-test; H2: the two groups will differ in post-test performance after pre-test scores are controlled.) RQ2: how consistently did learners in the two groups use their assigned application? RQ3: how did learners perceive the usefulness, difficulties and relative suitability of the two applications for learning English in Kazakhstan? (H3: the Oki/Aleem group will emphasise adaptivity and feedback more than the Duolingo group; H4: the Duolingo group will emphasise gamified engagement more than the Oki/Aleem group.) RQ4: what difficulties did learners report during weekly use and in their open-ended responses?

Methods and Materials. The design was a two-group pretest-posttest quasi-experiment with an added qualitative strand, so that the test results and the students' own comments could be read alongside each other. Over four weeks, two groups of undergraduates practiced with different mobile applications. Group A used Duolingo, a largely gamified application, and Group B used Oki/Aleem, an artificial-intelligence-based adaptive application developed in Kazakhstan. Participation was voluntary and the groups were not formed through random assignment, so the design is best understood as a non-equivalent-groups quasi-experiment. For this reason we examined, and report in the Results, whether the two groups were comparable before the intervention began. The primary and only objective outcome was the English achievement test. The perception scales had also been planned as outcomes, but their internal consistency proved too weak (see Results), so learners' experiences of the two applications are described from the open-ended answers instead. After giving informed consent, students were assigned in order of registration, alternating between the groups,

with the first eligible student allocated to Group A and the next to Group B, until both groups reached 60. Because no individual randomisation was carried out, the study is treated as quasi-experimental. Alternating allocation is systematic rather than random, and registration order could in principle correlate with characteristics such as eagerness or prior proficiency; alternation, however, distributes any such ordering trend evenly across the two conditions, and the baseline equivalence of the groups was verified empirically before the intervention.

The study was carried out at O. Zhanibekov South Kazakhstan Pedagogical University in Shymkent, Kazakhstan, with undergraduates of the programme 6B01703 “Foreign Language: Two Foreign Languages”, which prepares future teachers of foreign languages. Students from the first, second and third years took part on a voluntary basis, giving a total of 120 participants split evenly between the two groups (60 in each). Of these, 115 could be matched across the pre- and post-tests through their self-generated codes (57 in Group A and 58 in Group B); the remaining five participants were excluded from the paired analyses because the codes they had entered at pre-test and post-test did not correspond.

The sample leaned toward the lower undergraduate years, with 68 first-year, 33 second-year and 19 third-year students, and toward intermediate proficiency. On a self-report basis, 81 students placed themselves at level B1, 26 at B2 and 13 at C1, and none reported an A-level. Because proficiency was self-reported rather than tested, we did not treat it as an objective baseline and used the pre-test score for all group comparisons. Age and gender were not recorded, a limitation discussed below.

Students were recruited through their regular English classes and were informed about the purpose and duration of the study, the comparison of two mobile applications, and the pre-test, weekly monitoring form and post-test. Taking part was voluntary. Students who agreed signed an electronic consent form online, and nobody who declined proceeded to the pre-test or the pre-survey. Eligibility criteria were straightforward: at least 18 years of age, enrolment at undergraduate level, studying English as a foreign language, ownership of a smartphone with internet access, and willingness to remain in the study for all four weeks. Several kinds of data were removed before analysis: records of anyone under 18 or without consent, records with a missing pre-test or post-test, records of students who never used their assigned application, and duplicates.

Ethical approval came from O. Zhanibekov South Kazakhstan Pedagogical University, which classified the study as a minimal-risk educational intervention. Before the pre-test and pre-survey, every participant signed an electronic consent form. Students were told in plain terms what the study was about, what the four weeks would involve and how the testing would work. They also knew that participation was voluntary, that their answers were anonymous and confidential, and that they could leave at any moment without consequences for their coursework. No personal identifiers were collected. The four-to-six-character codes that students created themselves were used only to link their pre-test, weekly monitoring and post-test responses. The files were password-protected and accessible to the research team only, and nothing is reported here below the aggregate level.

Intervention and procedure. The timeline was straightforward. On 25-26 April 2026 students sat the pre-test and completed the pre-survey. From 27 April to 23 May they practised with their assigned application, and on 25-26 May they took the post-test and the post-survey. To keep the workload equal in both conditions, the instructions were identical: 15-20 minutes of practice a day, on at least five days a week, together with weekly evidence of use in the form of a screenshot or a progress record. The teacher’s role was limited to technical help. No teaching and no additional content were provided, so that any difference between the groups would derive from the applications themselves. Throughout the four weeks students continued their regular timetable, in which English was taught in two 50-minute lessons per week, and the application practice was assigned as part of their independent study workload. Both groups used the publicly available versions of the applications current during the intervention period (April–May 2026).

Access to the paid application was controlled so that cost would not separate the two groups during the study. All 60 students in Group B received one month of free access to Oki/Aleem and none of them paid anything during the study. The access ran for exactly the intervention period and covered the lessons, practice activities, feedback and progress-tracking features on which the study

relied. Even though access was free, some Group B participants still commented on subscription cost in their open-ended answers, which we read as a general perception of the application rather than as a barrier they had experienced. Every Friday students completed a short adherence form, and four weeks of these forms produced 480 weekly records.

English achievement test. Achievement in English was measured with a 20-item multiple-choice test, given in two parallel forms, one before the intervention and one after it. The items were developed by the research team to assess general English achievement at approximately the intermediate level. Each form contained six vocabulary items, six grammar items, five reading-comprehension items and three items on application-based language use; each item had four options and was scored 1 or 0, which put the ceiling at 20 points. The two forms used different item wording but an identical structure: the same number of items per language aspect was kept in both forms, and comparable difficulty was sought by matching the grammatical and lexical complexity of the items and the length of the reading passages across the forms. Parallel rather than identical forms were chosen deliberately, in order to limit the practice effect of encountering the same items twice. The forms did not undergo separate piloting or formal expert validation, a point acknowledged among the limitations; their functioning was therefore examined directly through internal-consistency analysis. Internal consistency was estimated with the Kuder-Richardson formula 20 (KR-20), the special case of coefficient alpha for dichotomously scored items [14]. On the post-test it reached an acceptable level (KR-20 = .731), whereas on the pre-test it was lower (KR-20 = .506). The pre-test scores therefore carry more measurement error, which we take into account when interpreting the gains.

The pre-survey gathered background information together with three five-point Likert scales, ranging from 1 (strongly disagree) to 5 (strongly agree), on initial motivation, digital readiness and attitude toward application-based learning, with four items each. These three scales showed poor internal consistency, with Cronbach's alpha values of .12, -.18 and -.01, so they were not used as composite measures, and baseline equivalence between the groups was established from the objective pre-test score instead. The post-survey had three parts. The first consisted of single questions on how often students had used the application, how much time they had spent and whether they had submitted weekly evidence. The second comprised seven self-constructed five-point scales intended to capture perceived learning gains, perceived adaptivity, gamified engagement, feedback quality, satisfaction, cognitive load and concerns, and intention to continue. The third consisted of three open-ended questions on the most useful feature, the main difficulty, and which application the student considered more suitable for learning English in Kazakhstan.

Cronbach's alpha [14, p. 299] for the seven perception scales was inadequate in every case, running from -.16 to .46. Even after the weakest item was dropped from each scale, no value climbed past the usual .70 threshold; the best we obtained was .51. These scales were therefore not analysed as composite measures and no scale-level conclusions were drawn from them; this outcome is reported as an instrument-validation finding and discussed among the limitations. Where individual items are mentioned, they are treated descriptively only.

Weekly monitoring. Each week a short form recorded how many days the application had been used, the average daily time, the skill the student had practised most, and one difficulty experienced that week. The original weekly records did not contain the participant code, and codes were reconstructed only for technical alignment of the files. For this reason, the weekly data are analysed at the group-by-week level, and no claims are made about individual learning trajectories or about links between an individual's usage and their achievement.

The quantitative data were analysed with descriptive statistics followed by a small set of inferential tests. Baseline equivalence between the groups was checked with a Welch's independent-samples t-test on the pre-test score. Change from pre-test to post-test was tested with paired-samples t-tests for the whole sample and within each group, and gain scores are reported with 95% confidence intervals – a precaution taken because the modest reliability of the pre-test inflates the uncertainty of change scores, and the difference between groups on the post-test, after adjusting for pre-test performance, was tested with a one-way analysis of covariance using the pre-test score as the covariate; the homogeneity of regression slopes was checked beforehand by testing the pre-test $\times$

group interaction. Effect sizes were reported as Cohen's d for the between-group comparison and as the standardised mean change for the paired comparisons. Reliability was estimated with the Kuder-Richardson formula 20 for the achievement test and Cronbach's α for the Likert scales, and the significance level was set at .05. Analyses were carried out in Python using the pandas, SciPy and statsmodels libraries. The three open-ended questions were analysed through inductive thematic analysis [15]: the responses were read repeatedly, coded for recurring ideas and grouped into themes, which were then tabulated by group and illustrated with representative quotations before being integrated with the quantitative results. The coding was carried out by the research team without an independent second coder, given the brevity and largely low-inference nature of the answers, single-coder inductive coding was judged acceptable for this exploratory strand, and the absence of double-coding and inter-rater agreement statistics is acknowledged among the limitations.

Results and discussion. Baseline comparability of the two groups. Because the groups were not formed by random assignment, we first checked whether they were comparable before the intervention. On the pre-test, the Duolingo group ($M = 12.68$, $SD = 3.15$; here and below M denotes the mean and SD the standard deviation) and the Oki/Aleem group ($M = 12.64$, $SD = 2.83$) performed almost identically, and the difference was not statistically significant, $t(113) = -0.12$, $p = .903$, $d = -0.02$ (t – t -test statistic, p – significance level, d – Cohen's d effect size). In other words, the two groups began at an equivalent level of English achievement, and their gains can be compared on that footing.

English achievement gains. Both groups improved markedly from the pre-test to the post-test. In the matched sample as a whole ($n = 115$), the mean score rose from 12.66 ($SD = 2.98$) to 15.63 ($SD = 3.32$), a large and highly significant gain of 2.97 points, 95% CI [2.65, 3.28] (CI – Confidence Interval), $t(114) = 18.58$, $p < .001$, standardised mean change 1.73. Each group showed the same pattern. Duolingo users gained 2.81 points, 95% CI [2.34, 3.28], and Oki/Aleem users 3.12, 95% CI [2.69, 3.55], in both cases at $p < .001$. These results support H1, which predicted significant improvement in both conditions. The corresponding indicators are presented in Table 1.

Table 1. Pre-test to post-test English achievement by group (matched sample).

Group	n	Pre-test M (SD)	Post-test M (SD)	Gain	t	p	d
Whole sample	115	12.66 (2.98)	15.63 (3.32)	+2.97	18.58	<.001	1.73
A: Duolingo	57	12.68 (3.15)	15.49 (3.50)	+2.81	11.92	<.001	1.58
B: Oki/Aleem	58	12.64 (2.83)	15.76 (3.16)	+3.12	14.45	<.001	1.90

Source: Compiled by the authors. Note: n – number of participants; M – mean; SD – standard deviation; t – t -test statistic; p – significance level; d – standardised mean change (paired-samples effect size).

To test whether the applications differed in their effect, we ran an analysis of covariance on the post-test scores with the pre-test score as the covariate. The assumption of homogeneity of regression slopes was met, as the pre-test $\times$ group interaction was not significant, $F(1, 111) = 0.00$, $p = .991$. The covariate strongly predicted post-test performance, $F(1, 112) = 315.03$, $p < .001$, partial $\eta^2 = .74$ ($b = 0.96$, where b denotes the unstandardised regression coefficient), whereas the effect of group was small and not significant, $F(1, 112) = 0.95$, $p = .332$, partial $\eta^2 = .008$, $b = 0.31$ in favour of Oki/Aleem, 95% CI [-0.32, 0.94]. After adjusting for pre-test performance, the estimated post-test means were 15.47 for Duolingo and 15.78 for Oki/Aleem. A difference between the applications was thus not detected in this sample, and H2, which anticipated one, was not supported.

Application use during the intervention. The weekly monitoring records, analysed at the group-by-week level, show that students in both groups used their application regularly and to a similar degree. On average, the Duolingo group practiced on 3.85 days per week ($SD = 1.16$) and the Oki/Aleem group on 3.69 days ($SD = 1.20$). Engagement was not flat across the month. In both groups the number of practice days climbed steadily from the first week to the fourth, which suggests that students settled into a routine rather than losing interest. The weekly distribution is shown in Table 2.

Table 2. Mean number of practice days per week, by group and week.

Group	Week 1	Week 2	Week 3	Week 4
A: Duolingo	2.63	3.62	4.37	4.78
B: Oki/Aleem	2.43	3.37	4.20	4.75

Source: Compiled by the authors.

Daily time on task was substantial in both conditions, and the reported skill focus was spread fairly evenly across vocabulary, grammar, listening, speaking and mixed practice. The post-survey pointed in the same direction. Most students reported using their application five days a week, if not daily, and nearly all reported submitting weekly evidence every week or almost every week. Adherence was therefore steady in both groups, which matters because it means that the similar achievement gains cannot be explained by unequal practice.

Reliability of the instruments. The achievement test itself behaved reasonably: KR-20 was .731 on the post-test, which is acceptable, but only .506 on the pre-test. The self-report scales did not perform well. The seven post-intervention perception scales returned Cronbach's alpha values between $-.16$ and $.46$, all below the $.70$ threshold, and removing the weakest item lifted the best value only to $.51$; the three pre-survey scales were no better ($.12$, $-.18$ and $-.01$). The items inside each scale simply did not hang together, so no scale could be trusted as a composite score; when examining how the two applications were perceived, we therefore turned to the open-ended answers instead of the ratings.

Learners' perceptions of the two applications. The three open-ended questions were analysed thematically, and the themes were tabulated by group. Themes were identified through inductive coding of the responses, and the reported frequencies are indicative. Although both applications produced similar measured gains, students described their value in very different terms. Asked about the most useful feature, Duolingo users wrote overwhelmingly about game-like motivation, including points, badges, leagues, streaks and reminders. Oki/Aleem users pointed instead to feedback and adaptivity, mentioning clear corrections, an artificial intelligence tutor they could question, speaking and pronunciation practice, lessons tuned to their level, and locally relevant content such as Kazakh cultural topics.

The difficulties students reported corresponded to these strengths. Duolingo users most often complained that the gamified design could become distracting or even stressful. Some wrote about pursuing points instead of learning, others about feeling pressured by the competition, and repetitive content and advertisements were also mentioned. With Oki/Aleem the complaints were mostly practical. Students mentioned the subscription price as a general concern, although access was free during the study, together with occasional technical glitches and the absence of any social or competitive dimension. On the question of suitability, Oki/Aleem was most often recommended for serious, exam-oriented study and for its local and cultural relevance, while Duolingo was praised for being enjoyable, motivating and easy to fit into daily life. A number of students concluded that the two tools are complementary. These patterns support H3 and H4. Duolingo was experienced primarily as a gamified, motivational tool and Oki/Aleem as an adaptive, feedback-rich and locally grounded one.

This study compared a gamified application (Duolingo) and an artificial-intelligence-based adaptive application (Oki/Aleem) within a single four-week intervention among Kazakhstani learners of English who are training to become foreign-language teachers. Three main findings emerged. Both groups improved substantially in English achievement, and once baseline scores were taken into account there was no significant difference between them. Both groups used their application consistently, and their practice increased across the four weeks. Although the measured gains were similar, the two applications were valued for very different reasons.

The first finding, a substantial improvement in both groups, supports H1. It is consistent with the wider literature, which indicates that mobile practice is associated with measurable language gains [1, p. 160]. For Duolingo there is direct evidence that completing its courses raises reading and listening proficiency [7, p. 263], and for adaptive instruction there is evidence of gains on difficult grammatical structures [10, p. 226]. The absence of a control group means that the gains cannot be attributed to the applications alone. Ordinary coursework and a testing effect almost certainly contributed, and the lower reliability of the pre-test adds noise to the change scores, so the effect should be read with the caution that meta-analysts have urged for this field [1, pp. 161-162]. The second finding is more informative. After controlling for pre-test performance, the two applications produced statistically indistinguishable outcomes, so H2 was not supported. This convergence mirrors comparative work in which Duolingo and classroom instruction brought beginners to similar levels [9, pp. 22-23], and it extends that pattern to a comparison of a gamified application with an adaptive one. It is also consistent with review evidence indicating that the effectiveness of gamified tools in foreign language learning is mixed rather than uniformly positive [6, p. 12], and that at university level the gamification effect is small and not statistically significant [5, p. 11]. Two different mechanisms, motivation through game design and personalisation through adaptive feedback, appear to produce comparable outcomes over a short intervention.

In answer to RQ2, learners in both groups practiced on a similar number of days each week, and in both groups practice rose steadily rather than declining. This is notable, since maintaining engagement in self-directed mobile study is a recognised difficulty [6, p. 2]. It is also in line with the finding that gains follow the regularity of practice more closely than the time invested [8, p. 24]. Because adherence was comparable, the similar achievement gains are easier to interpret. The groups did not differ in how much they practised, only in what they valued.

The qualitative analysis supported H3 and H4. Duolingo users described its value in terms of game-like motivation, which is the mechanism that the gamification literature associates with higher engagement and, through it, achievement [5, p. 8]. The same learners named the corresponding drawback, that of pursuing points rather than learning, a tension that reviewers of gamified learning have identified [6, p. 11]. Oki/Aleem users placed adaptivity and feedback first. These are the strengths that the literature attributes to artificial-intelligence tools, namely personalised, feedback-rich interaction that supports engagement, enjoyment and motivation [11, p. 6], although the evidence base as a whole remains young [12, p. 9]. The locally grounded content that several Oki/Aleem users valued corresponds to a recurring theme in Kazakhstani research, where relevance to the national context is treated as an asset of digital learning environments [3, p. 284]. Read together with the achievement data, these accounts suggest that, while no difference in test-score gains was detected between the two applications, they were experienced as serving different purposes. Duolingo functioned as a motivational, habit-forming tool and Oki/Aleem as a focused, adaptive tutor, and several learners concluded that the two are complementary.

A methodological finding requires separate comment. The self-constructed perception scales failed to reach acceptable internal consistency, even after a careful re-administration that produced the full range of responses, and were therefore set aside in favour of the open-ended data. Short, self-made Likert scales of three or four items are common in application-evaluation studies, yet their reliability is rarely scrutinised, and acquiescent responding can undermine it. As survey data are increasingly produced under conditions of inattentive or assisted responding, reporting reliability transparently and setting unreliable scales aside is part of responsible practice. In the present case the open-ended data recovered the distinction that the scales had been designed to capture.

Our participants were future foreign-language teachers, which gives the findings a second layer of meaning. For learning, they indicate that the choice between a gamified and an adaptive application is less a contest over effectiveness than a decision about purpose, since no advantage for either application was detected in this sample, and the two can be combined. For teacher preparation, the study underlines why digital competence belongs at the centre of foreign-language teacher education in Kazakhstan. Future teachers who have themselves compared a gamified and an adaptive

application are better placed to make informed pedagogical choices, a competence that research on Kazakhstani teacher training argues should be deliberately cultivated [3, p. 284].

Limitations. These conclusions come with several qualifications. The design was quasi-experimental. The groups were made up of volunteers, were not formed at random and came from a single university, so the findings may not generalise, even though the two groups were equivalent on the objective pre-test. Without a non-application control group, the gains cannot be attributed to the applications alone, and a four-week window can capture only short-term effects. Proficiency levels were self-reported, and the reliability of the pre-test was only modest; the test forms were developed by the research team without separate piloting or formal expert validation. The perception scales failed and had to be dropped. The open-ended answers, informative as they were, cannot be verified for provenance in the way an objective test can, and they were coded without an independent second coder. The weekly monitoring files contained no genuine participant codes, which confined that analysis to the group level. Age and gender were not collected. Future work should use larger, multi-institution samples, a control condition, longer interventions, validated instruments and properly linked process data.

Conclusion. Over four weeks, students using the gamified application (Duolingo) and the artificial-intelligence-based adaptive application (Oki/Aleem) showed significant gains in English achievement, and once baseline scores were controlled the gains were of similar size. Practice in both groups was steady and increased as the month progressed. The applications differed in how they were experienced. Learners saw Duolingo as a motivational, habit-forming tool, and Oki/Aleem as an adaptive tutor with detailed feedback and local relevance. The perception scales did not reach acceptable reliability and were set aside, and the students' own written accounts therefore provided the sounder basis for interpreting perceptions. For English learning, the practical question is therefore not which type of application works better – no difference between them was detected in this sample – but what a learner wants from it. For foreign-language teacher preparation in Kazakhstan, the implication is that future teachers should be given the opportunity to work with such tools alongside each other and to build their digital competence from that direct experience. Larger, longer and controlled studies with validated instruments are needed to confirm and extend these conclusions.

Acknowledgements. This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant No. AP26101671, “Development of methodology for formation of the future foreign language teacher’s professional competence in the context of digital adaptive learning”; scientific supervisor A. N. Zhorabekova).

REFERENCES

- 1 Burston, J., & Giannakou, K. (2022). MALL language learning outcomes: A comprehensive meta-analysis 1994-2019. *ReCALL*, 34(2), 147-168. DOI: <https://doi.org/10.1017/S0958344021000240>
- 2 Lei, X., Fathi, J., Noorbakhsh, S., & Rahimi, M. (2022). The impact of mobile-assisted language learning on English as a Foreign Language learners' vocabulary learning attitudes and self-regulatory capacity. *Frontiers in Psychology*, 13, 872922. DOI: <https://doi.org/10.3389/fpsyg.2022.872922>
- 3 Sarzhanova, G. B., Abdrakhmanova, S. T., Akbayeva, G. N., & Tleuzhanova, G. K. (2024). Digital learning environment in training future Kazakhstani English teachers. *Forum for Linguistic Studies*, 6(5), 280-286. DOI: <https://doi.org/10.30564/fls.v6i5.7039>
- 4 Szabó, F., & Kopinska, M. (2023). Gamification in foreign language teaching: A conceptual introduction. *Hungarian Educational Research Journal*, 13(3), 418-428. DOI: <https://doi.org/10.1556/063.2023.00202>
- 5 Lee, J.-Y., & Baek, M. (2023). Effects of gamification on students' English language proficiency: A meta-analysis on research in South Korea. *Sustainability*, 15(14), 11325. DOI: <https://doi.org/10.3390/su151411325>
- 6 Luo, Z. (2023). The effectiveness of gamified tools for foreign language learning (FLL): A systematic review. *Behavioral Sciences*, 13(4), 331. DOI: <https://doi.org/10.3390/bs13040331>

7 Jiang, X., Peters, R., Plonsky, L., & Pajak, B. (2024). The effectiveness of Duolingo English courses in developing reading and listening proficiency. *CALICO Journal*, 41(3), 249-272. DOI: <https://doi.org/10.1558/cj.26704>

8 Sudina, E., & Plonsky, L. (2024). The effects of frequency, duration, and intensity on L2 learning through Duolingo: A natural experiment. *Journal of Second Language Studies*, 7(1), 1-43. DOI: <https://doi.org/10.1075/jsls.00021.plo>

9 Kim, Y., Payant, C., Skalicky, S., & Namkung, Y. (2026). Comparing the effectiveness of Duolingo, classroom instruction, and classroom + Duolingo instruction conditions on beginner-level French language development. *Studies in Second Language Acquisition*, 1-28. DOI: <https://doi.org/10.1017/S0272263126101521>

10 Hadi, M. W., Suryadi, H., & Suparlan, S. (2026). The effectiveness of adaptive artificial intelligence (AI)-based learning in enhancing EFL learners' English structure mastery. *JOLLT Journal of Languages and Language Teaching*, 14(1), 223-235. DOI: <https://doi.org/10.33394/jollt.v14i1.17708>

11 Yuan, L., & Liu, X. (2025). The effect of artificial intelligence tools on EFL learners' engagement, enjoyment, and motivation. *Computers in Human Behavior*, 162, 108474. DOI: <https://doi.org/10.1016/j.chb.2024.108474>

12 AlTwijri, L., & Alghizzi, T. M. (2024). Investigating the integration of artificial intelligence in English as foreign language classes for enhancing learners' affective factors: A systematic review. *Heliyon. A Cell Press journal*, 10(10), e31053. DOI: <https://doi.org/10.1016/j.heliyon.2024.e31053>

13 Kurakbayeva, A., & Xembayeva, S. (2025). Enhancing professional abilities of university students through digital educational interventions: A study in Kazakhstani universities. *Frontiers in Education*, 9, 1478622. DOI: <https://doi.org/10.3389/educ.2024.1478622>

14 Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. DOI: <https://doi.org/10.1007/BF02310555>

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

Жорабекова А.¹, *Дуйсенова М.², Курмашева М.³, Есназар А.⁴

¹ М. Әуезов атындағы Оңтүстік Қазақстан университеті

^{2,3,4} Ө. Жәнібеков атындағы Оңтүстік Қазақстан педагогикалық университеті

^{1,2,3,4} Қазақстан, Шымкент

ҚАЗАҚСТАНДАҒЫ УНИВЕРСИТЕТ СТУДЕНТТЕРІНІҢ МОБИЛЬДІ АҒЫЛШЫН ТІЛІН ҮЙРЕНУІНДЕГІ ЖАСАНДЫ ИНТЕЛЛЕКТ НЕГІЗІНДЕГІ БЕЙІМДЕЛГІШ ҚОСЫМША МЕН ГЕЙМИФИКАЦИЯЛАНҒАН ҚОСЫМШАНЫ САЛЫСТЫРУ

Аңдатпа

Мобильді қосымшалар ағылшын тілін шет тілі ретінде үйренудің негізгі құралына айналды. Ал, нарықта екі тәсіл басым: геймификация және жасанды интеллект негізіндегі бейімделмелі оқыту. Бұл тәсілдерді тікелей салыстыратын зерттеулер сирек кездеседі, әсіресе Қазақстан жағдайында және елдің тілдік білім беру жүйесіне маман даярлау аясында болашақ шетел тілі мұғалімдерін даярлау саласында. Бұл зерттеуде геймификацияланған Duolingo қосымшасы мен Қазақстанда жасанды интеллект технологиялары негізінде әзірленген бейімделмелі Oki/Aleem қосымшасының тиімділігі төрт апталық кезең ішінде салыстырылды. Екі топтық квазиэксперименттік дизайн қолданылды: шетел тілі мұғалімдерін даярлау бағдарламасының 120 студенті екі қосымшаның бірімен жұмыс істеді; 115 студенттің деректері алдын ала және қорытынды тест нәтижелері бойынша сәйкестендірілді. Оқу жетістіктері зерттеу кезеңіне дейін және одан кейін өткізілген, 20 тапсырмадан тұратын параллель тестілер арқылы өлшенді; қосымшаларды пайдалану апталық сауалнама арқылы қадағаланды, ал үш ашық сұраққа берілген жауаптар тақырыптық тұрғыда талданды. Екі топ та бір ай ішінде көрсеткіштері айтарлықтай өсті; алдын ала тестілеу нәтижелерін ковариата ретінде ескере отырып жүргізілген ковариациялық талдау қосымшалардың тиімділігі арасында статистикалық маңызды айырмашылық анықталмады. Апта сайынғы оқу белсенділігі екі топта да сақталып, зерттеу кезеңінің соңына қарай өсті. Ашық сұрақтарға берілген жауаптарды талдау қатысушылардың зерттелген қосымшаларды қабылдауындағы айырмашылықтарды көрсетті: Duolingo ойын элементтері арқылы ынталандыруы, ал Oki/Aleem бейімделгіштігімен, кері байланысымен және жергілікті мазмұнымен бағаланды. Осылайша, екі қосымша әртүрлі механизмдерді қолданғанына қарамастан салыстырмалы оқу нәтижелерін көрсетті, бұл олардың арасындағы таңдауды тиімділік емес, білім беру

мақсаттары тұрғысынан қарастыруға мүмкіндік береді. Алынған нәтижелер Қазақстандағы шетел тілі мұғалімдерін даярлау бағдарламаларына геймификацияланған және бейімделмелі білім беру қосымшаларын салыстырмалы пайдалану тәжірибесін енгізудің орындылығын растайды.

Түйінді сөздер: жасанды интеллект, ағылшын тілі шет тілі ретінде, геймификация, Duolingo, Oki/Aleem, мобильді тіл үйрену, шетел тілі мұғалімдерін даярлау.

Жорабекова А.¹, *Дуйсенова М.², Курмашева М.³, Есназар А.⁴

¹ Южно-Казахстанский университет имени М. Ауэзова

^{2,3,4} Южно-Казахстанский педагогический университет имени О. Жанибекова

^{1,2,3,4} Казахстан, Шымкент

СРАВНЕНИЕ АДАПТИВНОГО ПРИЛОЖЕНИЯ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И ГЕЙМИФИЦИРОВАННОГО ПРИЛОЖЕНИЯ ДЛЯ МОБИЛЬНОГО ИЗУЧЕНИЯ АНГЛИЙСКОГО ЯЗЫКА СРЕДИ СТУДЕНТОВ УНИВЕРСИТЕТОВ КАЗАХСТАНА

Аннотация

Мобильные приложения стали центральным инструментом изучения английского языка как иностранного; на рынке доминируют два подхода – геймификация и адаптивное обучение на основе искусственного интеллекта. Прямые сравнительные исследования этих подходов встречаются редко, особенно в контексте Казахстана и подготовки будущих учителей иностранных языков для системы национального языкового образования. В настоящем исследовании сравнивалась эффективность геймифицированного приложения Duolingo и адаптивного приложения Oki/Aleem, разработанного в Казахстане на основе технологий искусственного интеллекта, в течение четырехнедельного периода. Использован квазиэкспериментальный дизайн с двумя группами: 120 студентов программы подготовки учителей иностранных языков работали с одним из двух приложений; данные 115 студентов сопоставлены по результатам пре- и посттеста. Учебные достижения измерялись с помощью параллельных тестов, состоявших из 20 заданий, которые проводились до и после вмешательства; еженедельная анкета использовалась для мониторинга использования приложений, а ответы на три открытых вопроса подвергались тематическому анализу. Обе группы значительно улучшили результаты в течение месяца; ковариационный анализ, выполненный с учетом результатов предварительного тестирования в качестве ковариаты, не выявил статистически значимых различий в эффективности между приложениями. Интенсивность еженедельной практики сохранялась в обеих группах и возрастала к концу периода вмешательства. Анализ ответов на открытые вопросы продемонстрировал различия в восприятии участниками исследуемых приложений. Duolingo ценился за игровую мотивацию, Oki/Aleem – за адаптивность, обратную связь и локальную релевантность. Таким образом, оба приложения обеспечили сопоставимые результаты обучения, несмотря на использование различных механизмов, что позволяет рассматривать выбор между ними скорее как вопрос образовательных целей, чем эффективности. Полученные результаты подтверждают целесообразность включения в программы подготовки учителей иностранных языков в Казахстане практики сравнительного использования геймифицированных и адаптивных образовательных приложений.

Ключевые слова: искусственный интеллект, английский язык как иностранный, геймификация, Duolingo, Oki/Aleem, мобильное изучение языка, подготовка учителей иностранных языков.

Received: 11.07.2026

Approved after peer review: 07.09.2026

Accepted for publication: 25.09.2026