Jurnal Pendidikan Progresif
e-ISSN: 2550-1313 | p-ISSN: 2087-9849
URL: https://jpp.fkip.unila.ac.id/index.php/jpp
Jounal's DOI: http://doi.org/10.23960/jpp

Vol. 16, No. 2, 2026
Article ID: 1429
Pages: 809-834

Two-Year Institutional Adoption of the EnglishScore App and Its Effects on EFL Proficiency and Engagement

Rizky Eka Prasetya[1,*], Adnan Mayof[2], & Achmad Syarif[1]
[1]Department of Secretarial Studies, Universitas Budi Luhur, Indonesia
[2]Department of English, Washington State University, United States

*Corresponding email: rizky.ekaprasetya@budiluhur.ac.id

Received: 09 December 2025 *Accepted: 14 April 2026* *Published: 21 May 2026*

Abstract: This study investigates the relationship between the two-year institutional adoption of the EnglishScore mobile assessment application and changes in proficiency and engagement among English as a Foreign Language (EFL) learners at Universitas Budi Luhur. It also explores learners' and instructors' perceptions of the app's integration into the EFL curriculum. A mixed-methods research design was employed. Quantitative data were collected from 120 EFL students through EnglishScore pre- and post-test results and a learner engagement questionnaire. Qualitative data were gathered through semi-structured interviews with 10 students and five instructors, as well as two focus group discussions. Quantitative data were analyzed using paired-samples t-tests and Pearson correlation analysis, while qualitative data were examined through thematic analysis. The results indicated statistically significant differences between pre-test and post-test proficiency scores in grammar, vocabulary, reading, and listening. Engagement levels were generally high, and a moderate-to-strong positive correlation ($r = 0.580$, $p < .001$) was found between engagement scores and proficiency gains. Qualitative findings suggested that learners perceived increased motivation, autonomy, and enjoyment when using the app. Instructors reported that the app was useful for diagnostic purposes, although limitations such as the absence of speaking and writing components and occasional technical issues were noted. The findings demonstrate a positive association between sustained use of the EnglishScore app, learner engagement, and observed proficiency gains over the two years. While the results suggest that the app may support language development, further research incorporating control groups or experimental designs is needed to establish causal relationships.

Keywords: mobile-assisted language learning, englishscore App, EFL proficiency, learner engagement, institutional integration.

Article's DOI: https://doi.org/10.23960/jpp.v16i2.pp809-834

INTRODUCTION

There have been significant changes in pedagogical practices concerning the use of mobile-assisted language learning tools in teaching English as a Foreign Language over the last ten years. With the increasing importance of English as a global language, digital pedagogy has been embraced by educational institutions in even non-native English speaking countries. One of these tools is the EnglishScore app, designed by the British Council, a mobile app for English learning and assessment.

Mobile-Assisted Language Learning involves the use of mobile technologies in language learning and teaching. Liu et al. (2024) stipulates that the use of mobile phones in educational settings is flexible, thus facilitating learning both in and out of the classroom. Recent studies (e.g. Tran et al. (2021); Kul (2023) have been concerned with the augmentation of educational opportunities that mobile applications bring and their influence on vocabulary learning and retention, listening skills enhancement, and learner autonomy. Research by Burston &

Giannakou (2022) found that MALL has a positive and significant impact on language learning outcomes, and even more so, when it is offered as part of the educational curriculum in educational institutions. While these outcomes are promising, there is a gap in studies, as there is little focus on long-term, MALL, institutional use of mobile technologies in language learning.

According to Okafor (2025), digital assessment technologies provide both formative and summative assessments. They provide instant feedback, monitor progress, and encourage self-directed learning. Tools such as Duolingo English Test, Versant by Pearson, and EnglishScore by the British Council have gained popularity for initial assessments, certifications, and feedback on class performance (Bouwer & Dirkx, 2023; Anton, 2021). Nevertheless, most studies focus on satisfaction and usability, leaving the educational impact largely unexplored. Zhang (2022) reported satisfaction with instant feedback on mobile tests. Their study, however, is limited to four weeks.

Currently, the British Council offers EnglishScore, a mobile diagnostic assessment tool, to evaluate learners' vocabulary and grammar, reading, and listening. Erzad (2024) and Birajdar (2020) explicated that the test is aligned with the Common European Framework of Reference (CEFR) and is increasingly recognized for both educational placement and employment purposes. Despite growing institutional interest in the app, Utami & Wahyudin (2022) and Apple et al. (2020) implied that peer-reviewed research examining its psychometric properties and pedagogical impact remains limited. While reports and technical documentation suggest that the test's validity and reliability are supported by large-scale pilot data, publicly accessible peer-reviewed validation studies are relatively scarce. Existing research has primarily focused on user perceptions rather than on the empirical evaluation of proficiency outcomes. For example, Sari et al. (2024) and Kurniati et al. (2025) reported positive learner attitudes toward EnglishScore in a Middle Eastern university context, but found no statistically significant short-term improvement in English proficiency. These findings highlight the need for more systematic, longitudinal research to examine the app's effectiveness and measurement properties in institutional settings.

Users can take CEFR-aligned diagnostic assessments on the EnglishScore platform, with instant results provided for various skills, including grammar, vocabulary, reading, and listening (Karatas & Fer. 2025). The combination of self-paced learning and accessible, adaptive design makes it attractive for self-directed learners and EFL educators aiming to facilitate standardized assessments and learning. In 2023, Universitas Budi Luhur adopted EnglishScore at the institutional level and embedded it into curricular and extracurricular teacher-training and assessment workflows. Over the following two years, the app served educators and learners not only as a diagnostic assessment tool but also as a means of deriving a motivational scaffolding approach for advanced language learning from big data analytics. EnglishScore, despite its popularity, has spawned little research to ascertain its impact on learning outcomes, especially on language proficiency and learner engagement.

Learner engagement is widely conceptualized as a multidimensional construct comprising behavioral, emotional, and cognitive components (Moorhouse & Wong (2020); Chen et al. (2020). Behavioral engagement refers to observable participation and effort in learning activities; emotional engagement involves interest, enjoyment, and affective responses; and cognitive engagement reflects strategic thinking, self-regulation, and investment in learning. Within EFL contexts, Zare et al. (2024) and Alam & Mohanty (2024) explained that these dimensions are closely linked to motivation and sustained participation in language tasks. To strengthen the

conceptual grounding of this study, it is important to map how specific features of the EnglishScore app may theoretically activate these engagement dimensions. First, the app's progress tracking and score reporting system, aligned with CEFR levels, may enhance cognitive engagement by encouraging self-monitoring and goal-setting behaviors (Jaelani et al. (2025) & Hidayati & Husna (2020). The ability to repeatedly attempt assessments and observe measurable improvement supports metacognitive regulation and strategic effort (Ramirez-Arellano. 2024). Second, the immediate feedback mechanism may reinforce behavioral engagement, as learners are prompted to reflect on errors and re-engage with content to improve subsequent performance (Ali et al. 2019). Third, the app's adaptive structure and score-based progression system may stimulate emotional engagement by creating a sense of challenge and achievement, resembling gamified learning environments (Maier & Klotz. 2022); Zolfaghari et al. 2025). From a motivational perspective, Self-Determination Theory (Chiu. 2022) provides a comprehensive framework for understanding how technology-supported learning environments foster intrinsic motivation and engagement. According to Self-determination theory (SDT), human motivation is fueled by the satisfaction of three basic psychological needs: autonomy (the need to feel volitional and self-directed), competence (the need to feel effective and capable), and relatedness (the need to feel connected and valued by others) (Autin et al. (2022). When individuals perceive that an activity satisfies these needs, they are more likely to engage authentically—behaviorally (through persistence and effort), emotionally (through enthusiasm and positive affect), and cognitively (through deep processing and metacognition) (Champ et al. (2023). Mobile language learning applications like EnglishScore can potentially satisfy these basic psychological needs through specific design features. Autonomy is supported through self-paced learning, choice of exercises, and flexible access, allowing learners to direct their own learning pathway. Competence is fostered through scaffolded difficulty levels, immediate corrective feedback, and progress visualization (e.g., achievement indicators, score tracking, CEFR level placement). Relatedness may be promoted through leaderboards, comparison with peers, or teacher-facilitated review and discussion of results. Consequently, well-designed mobile language apps create conditions favorable for the emergence of multidimensional engagement: behavioral (on-task participation), emotional (positive affect and interest), and cognitive (deep thinking and metacognition) (Li & Contreras. 2024) and Shadiev et al. (2017).

Although there has been an increase in the literature on mobile-assisted language learning (MALL) and digital assessment, critical gaps remain. For one, there are very few longitudinal studies investigating the use of mobile assessment tools in institutional EFL settings. Most of the research done in this area is studies that are very short in time (short-term studies) or lack an institutional focus and instead concentrate on individual learners. For another, there are very few studies that focus on how the EnglishScore app, despite its popularity, has been very influential in language learning (even pedagogically), as there are virtually no studies focusing on EnglishScore that are empirical, making the majority of the information available promotional in nature or not even reviewed or peer-reviewed. This research aims to assess the mobile EnglishScore app in light of empirical gaps. This research has captured and crystallized both student engagement with the EnglishScore app and EFL proficiency within an institutional, formal structure over two years, thereby elucidating the empirical evidence of the gaps. This research specifically focuses on the gaps to create the following research questions:

1. What has been the impact of the two-year institutional adoption of the EnglishScore app on the English language proficiency of students studying English as a foreign language?
2. How do students studying English as a foreign language and their instructors perceive the use of the EnglishScore app within the institution?
3. In what way has the use of the EnglishScore app contributed to student engagement within the English as a foreign language (EFL) classroom?

METHOD

Participants

The study was conducted at Universitas Budi Luhur, where EnglishScore was mandated as part of the EFL curriculum during the two-year academic period (2022-2024). Because all students enrolled in the relevant EFL courses were required to use the app, the sampling approach is best described as a convenience sample drawn from an accessible population of mandatory app users. During the two-year implementation period, approximately 210 students across all academic programs were enrolled in EFL courses that required the use of EnglishScore. All 210 students were required to complete EnglishScore pre-tests at the beginning of Year 1 and post-tests at the end of Year 2 as part of standard institutional assessment procedures.

A total of 210 students were originally enrolled in the study, but only those with complete data were included. To be eligible, participants had to have complete pre-test proficiency data from Year 1, complete post-test proficiency data from Year 2, and complete responses to the learner engagement questionnaire administered after Year 2. Based on these criteria, 90 students were excluded due to incomplete data: 35 who dropped out during the two years, 28 who missed the pre-test, 19 who missed the post-test, and 8 who did not complete the questionnaire. As a result, the final sample consisted of 120 students, representing 57.1% of the initial enrollment. All included participants had complete data across the three collection points, allowing the researchers to conduct paired comparisons and correlation analyses without missing data issues.

Table 1. Universitas budi luhur student participants (N = 120)

Category	Details	Number of Students	Percentage (%)
Duration of App Use	2 Academic Years	120	100%
Academic Program	Faculty of Social and Political Science	75	62.50%
	Faculty of Engineering	18	15.00%
	Faculty of Economics & Business	15	12.50%
	Faculty of Communication Sciences	7	5.80%
	Faculty of Information Technology	5	4.20%
Year of Study	Year 1–2	58	48.3%
	Year 3–4	62	51.7%
Gender	Female	72	60%
	Male	48	40%
Age Group	18–20 years	50	41.7%
	21–23 years	58	48.3%
	24 years and above	12	10%

CEFR Proficiency Level	Intermediate (B1)	68	56.7%
	Upper-Intermediate (B2)	52	43.3%
Primary Device Used	Smartphone	98	81.7%
	Laptop/PC	22	18.3%
Internet Access	Personal Wi-Fi or Mobile Data	100	83.3%
	Campus Wi-Fi only	20	16.7%

Table 2. Universitas budi luhur instructor (lecturer) participants (N = 10)

Category	Details	Number of Instructors	Percentage (%)
Duration of App Use	2 Academic Years	10	100%
Role in Implementation	Course Implementers	6	60%
	Supervisors/Program Coordinators	4	40%
Gender	Female	6	60%
	Male	4	40%
Years of Teaching Experience	1–5 years	3	30%
	6–10 years	5	50%
	More than 10 years	2	20%
Academic Rank	Lecturer	7	70%
	Assistant Professor	3	30%
Field of Expertise	English Language Teaching (ELT)	8	80%
	Applied Linguistics or Curriculum	2	20%
Familiarity with EdTech	Intermediate	6	60%
	Advanced	4	40%

Table 3. Sampling procedure and data completeness screening

Item	Number	Percentage
Total students enrolled in EFL courses (2022–2024)	210	100%
Students with complete pre-test data	182	86.7%
Students with complete post-test data	165	78.6%
Students with a complete engagement questionnaire	157	74.8%
Students with complete data across all three measures	120	57.1%
Exclusion reasons:		
Program dropout	35	16.7%
Missed pre-test	28	13.3%
Missed post-test	19	9.0%
Incomplete questionnaire	8	3.8%
Total excluded (incomplete data)	90	42.9%
Final sample (N)	120	57.1%

Research Design

This study employed a mixed-methods convergent parallel design that combines quantitative and qualitative approaches to understand the two-year institutional adoption of the EnglishScore app. The quantitative strand

examined changes in EFL learners' proficiency and engagement over time, while the qualitative strand explored learners' and instructors' perceptions. Both types of data were collected simultaneously and analyzed separately before being combined in the interpretive phase. This approach enabled a comprehensive understanding of the app's impact on both user experience and measurable learning outcomes.

The study was conducted over a two-year institutional implementation period. At the beginning of the first academic year, baseline proficiency data were obtained through EnglishScore pre-test results. Throughout the two years, students continued to use the EnglishScore app as part of their institutional EFL curriculum. At the end of the second academic year, post-test proficiency scores were collected from the app's institutional dashboard to measure changes over time. In addition, the learner engagement questionnaire was administered after sustained exposure to the application to assess behavioral, emotional, and cognitive engagement.

Simultaneously, qualitative data were collected through semi-structured interviews with selected students and instructors, as well as focus group discussions with students. These sessions were conducted after the two-year implementation period to ensure that participants could reflect on their extended experience using the app. Quantitative and qualitative data were analyzed independently according to their respective analytical frameworks. Finally, findings from both strands were compared and integrated during the interpretation stage to identify convergences, complementarities, and potential divergences between measured outcomes and participant perceptions.

Data Collection Procedures

A combination of quantitative and qualitative approaches was employed to understand the impact of the EnglishScore app on participants' EFL proficiency and engagement. Quantitative data were collected using two primary sources:

1. The app's institutional dashboard, which provided access to EnglishScore proficiency reports and enabled the retrieval of participants' pre- and post-test results across the two years. Standardized scores for grammar, vocabulary, reading, and listening were provided and aligned to CEFR levels.
2. An engagement questionnaire was administered post-study to measure behavioral, emotional, and cognitive engagement using a 5-point Likert scale.

Qualitative data were collected to complement the quantitative findings. Purposeful sampling was used to conduct semi-structured interviews with 10 students and five instructors who were prolific users of the app. These interviews focused on the app's perceived usability, its perceived integration into the curriculum, and the perceived benefits and challenges of incorporating it into instruction. To gain further insight into themes that emerged from the interviews (e.g., motivation, instant feedback, overall experience of mobile assessment in the EFL classroom), selected students were invited to participate in focus group discussions.

This study employed a single-group, quasi-experimental longitudinal design without a comparison or control condition. All participants used EnglishScore throughout the two years; no control group of non-users was available for comparison. Consequently, proficiency gains cannot be attributed solely to app use, as multiple confounding variables may have contributed to observed improvements, including: (1) academic maturation: students naturally develop in language proficiency over two years through concurrent enrollment in regular English courses and general academic development; (2) concurrent instruction: participants received instruction in other English courses beyond EnglishScore use, providing

alternative explanations for proficiency gains; (3) self-selection effects: students who consistently used the app may have differed in motivation, self-efficacy, or study habits compared to less-engaged users; and (4) environmental factors: access to English media, private tutoring, self-study initiatives, and socioeconomic resources may have varied among participants. Despite these limitations, the study provides valuable institutional-level evidence on the patterns and magnitude of association between app use and proficiency in a real-world educational setting, informing practice and guiding future research designs.

Instruments

The first instrument used in this study was the EnglishScore mobile assessment application developed by the British Council. The app assesses learners' English proficiency in four domains: grammar, vocabulary, reading, and listening. The assessment is aligned with the Common European Framework of Reference for Languages (CEFR), and scores are reported on a standardized scale ranging from 0 to 100. For this study, matched pre-test (Year 1) and post-test (Year 2) scores were extracted from institutional records. Only participants with complete data across both time points were included in the analysis. The app employs adaptive testing mechanisms and automated scoring, ensuring consistency across administrations. Since EnglishScore is externally standardized, the researcher did not conduct internal reliability testing; however, institutional documentation reports acceptable reliability coefficients from pilot validation studies.

The second instrument was a learner engagement questionnaire adapted primarily from the Student Engagement Instrument (SEI) developed by Appleton et al. (2006), supplemented by constructs from contemporary engagement literature in mobile-assisted language learning contexts. The instrument was modified to reflect EFL-specific classroom engagement and technology-mediated learning. The engagement questionnaire was designed to measure three dimensions of student engagement (behavioral, emotional, and cognitive) as theoretically grounded in Self-Determination Theory (Shen et al. (2024). The instrument was adapted primarily from the Student Engagement Instrument (SEI) developed by T. Li et al. (2024), supplemented by constructs from contemporary engagement literature She et al. (2023) Moreover, intentionally aligned with SDT principles.

Table 4. Construct, indicators, and sample items of the learner engagement questionnaire

Dimension	Indicators	Sample Items	Number of Items
Behavioral Engagement	• Participation in EnglishScore activities • Completion of assigned tasks • Effort in improving test performance • Persistence in retaking assessments • Active involvement during class discussions	• "I try hard to improve my EnglishScore results." • "I complete EnglishScore-related tasks on time."	11
Emotional Engagement	• Enjoyment of using EnglishScore • Interest in improving scores • Feelings toward digital assessment • Sense of achievement • Positive classroom emotions	• "I enjoy using the EnglishScore app." • "I feel proud when my EnglishScore level improves."	10
Cognitive Engagement	• Self-monitoring of progress • Strategic learning behaviors •	• "I analyze my mistakes after receiving my	9

	Reflection on mistakes • Goal-setting behavior • Deep learning orientation	EnglishScore results." • "I set personal goals to improve my English proficiency."	
Total Items	—	—	**15**

All items were measured using a 5-point Likert scale (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree). Content validity was assessed using Aiken's V coefficient to quantify expert agreement.

Expert Selection Criteria. Three experts were selected based on the following criteria: (1) Doctoral degree in Applied Linguistics or ELT; (2) Minimum 5 years of experience in language assessment; (3) Prior publication in peer-reviewe journals on engagement or educational technology

Validation Procedure

Each expert evaluated items on: (1) Item relevance to construct; (2) Clarity of wording; (3) Cultural and contextual appropriateness; (4) Alignment with EFL engagement theory. Items were rated on a 5-point relevance scale. Aiken's V values ranged from 0.83 to 0.94, exceeding the acceptable threshold of 0.80. Revisions Made: (1) Two items were reworded for clarity; (2) an item was removed due to redundancy; (3) Minor linguistic adjustments were made for contextual appropriateness. The third instrument was a semi-structured interview protocol developed to gather insights from students and instructors regarding their perceptions of the app's usability, integration, and impact. This protocol was piloted and refined based on participant feedback to ensure clarity, relevance, and completeness.

Internal Consistency and Reliability

In addition to content validity established through expert evaluation using Aiken's V, the internal consistency reliability of the engagement questionnaire was assessed using responses from the 120 study participants. Cronbach's alpha (á) was calculated for each engagement subscale and for the overall scale to determine how consistently the items measured the intended constructs.

Table 5. Internal consistency reliability of engagement subscales

Engagement Dimension	Number of Items	Cronbach's Alpha (α)	Interpretation
Behavioral Engagement	11	0.82	Good
Emotional Engagement	10	0.79	Acceptable
Cognitive Engagement	9	0.76	Acceptable
Overall Engagement Scale	30	0.85	Good

Note: An alpha coefficient of .70 or higher indicates acceptable internal consistency, while .80 or higher indicates good internal consistency. All subscales exceeded the .70 threshold, indicating that the instrument demonstrated adequate reliability for this study.

Among the three subscales, Behavioral Engagement showed the highest internal consistency (á = .82), indicating that the 11 items related to effort, persistence, participation, and task completion were strongly interrelated. Emotional Engagement (á = .79) and Cognitive Engagement (á = .76) also demonstrated acceptable internal consistency, although at slightly lower levels. The overall 30-item engagement scale yielded a Cronbach's alpha of .85, showing good internal consistency and suggesting that the instrument reliably measured multidimensional learner engagement. Taken together with the content validity results from expert review (Aiken's V = 0.83–0.94), these findings support the validity and reliability of the

instrument for assessing engagement in this EFL mobile learning context.

Data Analysis

Quantitative Data Analysis

Descriptive statistics (means, standard deviations, minimum, maximum, and ranges) were calculated to summarize EnglishScore proficiency scores across four skill domains (grammar, vocabulary, reading, and listening) at pre-test (Year 1) and post-test (Year 2). To determine whether statistically significant improvements in proficiency occurred over the two years, paired-samples t-tests were conducted to compare pre-test and post-test proficiency scores for each skill domain. This test was appropriate because the same 120 students provided both pre- and post-test data, and proficiency scores are continuous variables. Paired t-tests were conducted separately for: (1) overall proficiency score, (2) grammar domain, (3) vocabulary domain, (4) reading domain, and (5) listening domain. Effect sizes (Cohen's d) were calculated for all comparisons to quantify the magnitude of improvement independent of sample size. Statistical significance was set at á = .05 (two-tailed).

Analysis: Descriptive statistics were calculated for each engagement subscale (Behavioral, Emotional, and Cognitive) using responses from all 120 participants. Engagement subscale scores were computed as the mean of constituent items (range 1–5) for interpretability. Correlation analysis (Pearson r) was conducted to examine bivariate relationships between each engagement subscale and proficiency gain (post-test minus pre-test overall proficiency score). Pearson correlations were used because both engagement and proficiency gain are continuous variables. Three bivariate correlations were calculated: (1) Behavioral Engagement vs. Proficiency Gain, (2) Emotional Engagement vs. Proficiency Gain, and (3) Cognitive Engagement vs. Proficiency Gain. All quantitative analyses were conducted using SPSS Version 27.0 (IBM Corp., 2020), with statistical significance set at á = .05 (two-tailed).

To examine the unique contribution of each engagement dimension to students' proficiency improvement, a multiple regression analysis was conducted with Proficiency Gain as the dependent variable. Proficiency Gain was calculated by subtracting the pre-test proficiency score from the post-test proficiency score, thereby representing the magnitude of improvement over the two years. The score ranged from -25 to +50 points. Four predictors were entered into the model: Pre-Test Proficiency, Behavioral Engagement, Emotional Engagement, and Cognitive Engagement. Pre-Test Proficiency was included as a covariate to control for baseline differences, potential floor or ceiling effects, and regression to the mean. The three engagement variables were represented by the mean scores of their respective subscales, each measured on a 1- to 5-point scale. The regression model was specified as follows:

Proficiency Gain = β_0 + β_1 (Pre-Test Proficieancy + β_2 (Behavioral Engagent) + β_3 (Emotional Engagement) + β_4 (Cognitive Engagement) + $_E$

where β_0 represents the intercept, β_1 to β_4 represent the unstandardized regression coefficients, and εrepresents the error term. All predictors were entered simultaneously using the Enter method, rather than a stepwise procedure, because all variables were considered theoretically relevant. For each predictor, the analysis produced unstandardized coefficients (B), standardized coefficients (*B*), standard errors (SE), t-values, p-values, and 95% confidence intervals.

Before interpreting the model, the assumptions of multiple regression were examined. Linearity was assessed through scatterplots comparing each predictor with

proficiency gain, and the relationships appeared approximately linear. Independence of observations was assumed because the data were obtained from independent participants and no repeated-measures or nested structure was present. Homoscedasticity was evaluated by inspecting a scatterplot of standardized residuals against standardized predicted values, which suggested reasonably even variance. Normality of residuals was examined using a histogram, a normal Q-Q plot, and the Shapiro-Wilk test (p = [insert value]) . If the Shapiro-Wilk result was non-significant, residuals were considered approximately normally distributed; if significant, any minor deviation from normality was considered acceptable given the sample size (N = 120), as multiple regression is generally robust to modest violations of this assumption. Multicollinearity was assessed using Variance Inflation Factors (VIF), and all VIF values were below 5.0 (range = [insert range]), indicating that multicollinearity was not problematic.

The overall model was evaluated using the F-statistic, R^2, adjusted , and the standard error of the estimate. The unstandardized coefficient (*B*) for each predictor was interpreted as the expected change in proficiency associated with a 1-unit increase in that predictor, while holding all other variables constant. The standardized coefficient (*B*) was used to compare the relative strength of predictors measured on different scales. All analyses were conducted using SPSS version 27.0, with statistical significance set at á = .05 (two-tailed).

Table 6. Variables included in the multiple regression model

Variable	Role	Description	Scale
Proficiency Gain	Dependent variable	Post-test proficiency score minus pre-test proficiency score	Continuous (-25 to +50)
Pre-Test Proficiency	Predictor/covariate	Baseline proficiency score measured in Year 1	Continuous
Behavioral Engagement	Predictor	Mean score on the Behavioral Engagement subscale	Continuous (1–5)
Emotional Engagement	Predictor	Mean score on the Emotional Engagement subscale	Continuous (1–5)
Cognitive Engagement	Predictor	Mean score on the Cognitive Engagement subscale	Continuous (1–5)

Qualitative Data Analysis

Qualitative data were collected through semi-structured interviews and focus group discussions involving 10 students, 5 instructors, and 2 student focus groups consisting of 8–10 participants each. All sessions were audio-recorded with participants' consent, yielding approximately 18 hours of recordings. The recordings were transcribed verbatim by a professional transcriber and subsequently verified by the research team by comparing them with the original audio files. Any discrepancies were corrected to ensure transcription accuracy. The final transcript set comprised approximately 85 single-spaced pages, with only minimal editing to preserve confidentiality.

The qualitative data were analyzed using thematic analysis following Braun and Clarke (2006), supported by the interactive model of Miles, Huberman, and Saldaña (2014), which emphasizes data condensation, data display, and conclusion drawing or verification. The analysis began with repeated readings of the transcripts to become familiar with the data, accompanied by memo writing to capture initial impressions and emerging analytical ideas. Meaningful units of text relevant to the research questions were then identified and coded using a hybrid approach that

combined deductive and inductive coding. Deductive codes were derived from the study's theoretical framework, including engagement dimensions, Self-Determination Theory constructs, and implementation-related factors, whereas inductive codes emerged from participants' own language and experiences. A codebook was developed to define code names, inclusion criteria, and exclusion criteria, and it was revised iteratively throughout the analysis. Weekly research team meetings were conducted to review coded segments, refine code definitions, and ensure consistency in code application.

Following coding, the data were organized into matrices to compare participant groups (students and instructors) and data sources (interviews and focus groups). This process facilitated the identification of patterns, convergences, and divergences across cases. Codes were subsequently clustered into broader categories and developed into higher-order themes based on conceptual similarity and analytical relevance. Themes were evaluated according to their prevalence across participants, the intensity and specificity with which they were discussed, and their explanatory value in relation to the quantitative findings. In the final stage, themes were refined by comparing them with the full dataset, considering alternative interpretations, and examining deviant cases to ensure that the findings were grounded in the full range of participant perspectives. All qualitative data management and coding procedures were conducted using NVivo 12 (QSR International, 2018).

To enhance trustworthiness, several rigor strategies were employed. Data-source triangulation was achieved by collecting data from both students and instructors and by drawing on interviews and focus group discussions. Method triangulation was also applied by comparing qualitative findings with quantitative results during the mixed-methods synthesis stage. Peer debriefing was carried out through weekly team discussions of coding decisions, theme development, and interpretation. An audit trail was maintained throughout the analysis, including analytical memos, codebook revisions, decision logs, and thematic matrices. In addition, verbatim quotations were used to preserve participants' voices, and preliminary findings were shared with a subset of participants for member checking. These participants confirmed that the identified themes accurately reflected their experiences.

During the integration stage, qualitative and quantitative findings were compared to identify points of convergence and divergence. This process enabled the qualitative findings to provide contextual explanation for the quantitative results, particularly in clarifying how themes such as ownership of learning and self-motivation helped explain the stronger predictive role of behavioral engagement in language proficiency outcomes.

■ RESULT AND DISCUSSION

Preliminary Analyses and Instrument Reliability

Prior to testing primary research questions, preliminary analyses verified the reliability and appropriateness of measurement instruments. Cronbach's alpha coefficients for the engagement subscales were: Behavioral Engagement $\alpha = .82$ (11 items), Emotional Engagement $\alpha = .79$ (10 items), and Cognitive Engagement $\alpha = .76$ (9 items). All values exceeded the .70 threshold for acceptable internal consistency, indicating that the engagement measure was reliable for use with this sample. The overall engagement scale (30 items) demonstrated good internal consistency (á = .85).

Research Question 1: What has been the impact of the two-year institutional adoption of the EnglishScore app on the English language proficiency of students studying English as a foreign language?

To determine the EnglishScore app's impact on students' English language skills, we

collected pre- and post-app scores over two years from 120 EFL students. These scores spanned the app's four skill domains: grammar, vocabulary, reading, and listening, and corresponded to the CEFR levels. Each student took a diagnostic at the beginning of EnglishScore (Time 1) and at the end of the second academic year (Time 2).

Table 7. Descriptive statistics for englishscore proficiency scores (N = 120)

Skill Area	Pre-Test Mean (SD)	Post-Test Mean (SD)	Mean Difference
Grammar	62.1 (9.8)	71.3 (8.7)	+9.2
Vocabulary	64.7 (10.1)	74.2 (9.2)	+9.5
Reading	66.3 (11.4)	75.0 (10.3)	+8.7
Listening	60.9 (10.5)	70.6 (9.8)	+9.7
Overall	63.5 (10.5)	72.8 (9.5)	+9.3

Table 7 indicates a consistent upward trend across all four measured skill areas over the two years. Rather than merely reflecting numerical increases, the pattern of gains suggests differentiated development across skill domains. Notably, listening demonstrated the largest mean increase (+9.7), followed closely by vocabulary (+9.5). One possible explanation for the relatively stronger gains in listening may be the app's format, which emphasizes short, structured listening tasks delivered in a mobile environment. Repeated exposure to audio input and immediate feedback mechanisms may have supported incremental improvements in receptive processing skills. Similarly, vocabulary gains may reflect the app's structured item-based testing format, which reinforces lexical recognition and recall through repeated engagement.

In contrast, while reading and grammar also improved substantially, their slightly lower gains may reflect the broader influence of classroom instruction and academic coursework, rather than the application itself. Since reading tasks in university settings are often embedded in discipline-specific content, improvement in this domain may be shaped by multiple contextual factors. Importantly, although the observed increases in scores suggest meaningful developme over time, the one-group pretest–posttest design does not permit causal attribution. The improvements recorded over two academic years may also be influenced by maturation effects, exposure to other English courses, increased academic literacy demands, extracurricular English input (e.g., media consumption), or general cognitive development. Therefore, the results should be interpreted as indicating an association between sustained exposure to the EnglishScore application and observed proficiency gains, rather than definitive evidence that the app alone caused the improvement. Future research employing control groups or quasi-experimental designs would be necessary to isolate the application's specific contribution.

Table 8. Paired samples t-test results

Skill Area	Mean Difference	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Grammar	9.20	8.45	0.84	7.55	10.85	11.43	119	.000
Vocabulary	9.50	8.13	0.78	7.96	11.04	12.18	119	.000
Reading	8.70	9.00	0.81	7.10	10.30	10.75	119	.000
Listening	9.70	8.50	0.81	8.10	11.30	12.02	119	.000
Overall	9.30	8.27	0.73	7.86	10.74	12.84	119	.000

Table 8 presents the results of the Paired Samples t-Test. The entries in the Skill Area column contain the means, which are in the column to the right. The next column to the right contains the standard deviations. The next column over contains the differences in standard errors. The next column to the right contains confidence intervals for the differences. The lower and upper limits for the differences are in the last two columns to the right. In the last two columns, art- statistics and two-tailed significance statistics. As indicated in the table, the paired-sample t-test showed that all score changes were statistically significant. The lowest mean score change was 8.70, and the highest mean score change was 9.797, with p-values < .001. Given that all values have confidence intervals that do not include 0, we conclude that English influences positive change in learners' English proficiency over two years.

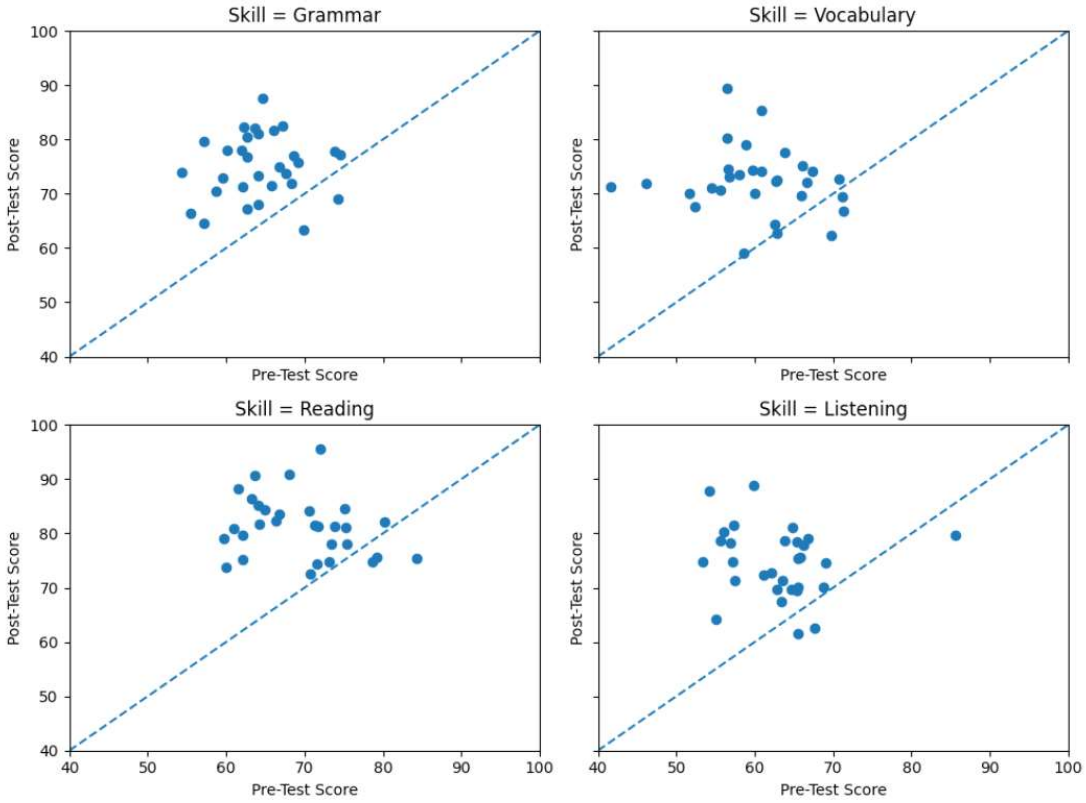

Figure 1. Scatter plot distribution of pre-test and post-test englishscore proficiency scores (N = 120) across four language skills

Scatter plots comparing individual pre-test and post-test scores across four skill areas: Grammar, Vocabulary, Reading, and Listening. Each point represents a learner's paired scores, while the diagonal reference line (y = x) indicates equal pre-test and post-test performance. Points located above the diagonal line reflect improvement, whereas points below the line indicate a decline. Across all four skills, the majority of data points lie above the diagonal reference line, indicating that most learners achieved higher post-test scores than their pre-test performance. This upward distribution is particularly noticeable in Listening and Vocabulary, where the dispersion of points suggests consistent gains among participants. Grammar and Reading also show clear positive shifts, although with slightly more variability in score changes. The clustering of points above the diagonal line across skill areas supports the

statistical findings of significant pre- to post-test differences. However, while the scatter plots illustrate a general pattern of improvement over the two years, the figure does not imply causality. The visual trend indicates that proficiency gains were observed during the institutional implementation of the EnglishScore app, but other contextual and developmental factors may also have contributed to the changes.

Research Question 2: How do students studying English as a foreign language and their instructors perceive the use of the EnglishScore app within the institution?

At the conclusion of the second year, all 120 students received the modified version of the Student Engagement Instrument (SEI), which included 15 items on a modified Likert 5-point scale (1 equivalent to Strongly Disagree and 5 to Agree Strongly) to measure the three components of engagement:

Behavioral (e.g., effort, participation)
Emotional (e.g., interest, enjoyment)
Cognitive (e.g., self-regulation, strategic thinking)

Although overall engagement levels were high, a notable pattern emerged across the three dimensions. Behavioral (M = 4.12) and Emotional Engagement (M = 4.08) were higher than Cognitive Engagement (M = 3.94), suggesting a differentiated engagement profile rather than uniformly strong involvement across

Table 9. Descriptive statistics – student engagement scores (N = 120)

	Likert Scale							
Engagement Dimension	**1**	**2**	**3**	**4**	**5**	**Mean**	**S.D**	**Interpretation**
Behavioral Engagement	2	4	18	56	40	4.12	0.54	High
Emotional Engagement	3	6	21	54	36	4.08	0.61	High
Cognitive Engagement	4	8	28	50	30	3.94	0.66	Moderate-High
Overall Engagement	3	6	22	53	36	4.05	0.60	High

Note: Engagement is interpreted as high to the degree that mean scores exceed 4.0

domains. While students demonstrated active participation in EnglishScore-related tasks and reported positive affective responses toward the application, their level of deeper cognitive investment such as strategic thinking, metacognitive reflection, and goal-oriented self-regulation was comparatively lower. This distinction is conceptually meaningful. The design features of the EnglishScore app, including structured diagnostic tasks, automated scoring, and immediate feedback, may effectively promote task completion and positive emotional responses. However, such features may provide limited scaffolding for higher-order cognitive processes that require extended reflection, problem-solving, or strategy development. In other words, the app appears to facilitate participation and motivation but may be less explicitly designed to cultivate deeper cognitive engagement. This finding aligns with engagement theory, which emphasizes that technological tools can successfully stimulate behavioral and emotional involvement without necessarily fostering sustained cognitive processing unless intentionally structured to do so. Therefore, while the institutional integration of EnglishScore is associated with high overall engagement, the comparatively lower cognitive dimension suggests an area for pedagogical enhancement, particularly through complementary instructional strategies that promote strategic learning and metacognitive awareness.

The correlation analysis reveals several important patterns regarding students' perceptions of the institutional use of the EnglishScore app. Overall engagement demonstrated a moderate-to-strong positive correlation with proficiency gain ($r = .580$, $p < .001$), indicating that students who reported higher levels of engagement tended to

Table 10. Correlation matrix between engagement and proficiency gain (N = 120)

Variables	1	2	3	4	5	6
1. Overall Engagement	1.00					
2. Behavioral Engagement	0.88**	1.00				
3. Emotional Engagement	0.85**	0.72**	1.00			
4. Cognitive Engagement	0.83**	0.69**	0.71**	1.00		
5. Pre-Test Proficiency	-0.21*	-0.18	-0.15	-0.12	1.00	
6. Proficiency Gain	0.580**	0.552**	0.497**	0.421**	-0.312**	1.00

*N = 120 Pearson correlation (2-tailed). * p < .05 ** p < .01

Table 11. Pearson correlation between predictors and proficiency gain (N = 120)

Predictor	Pearson's Correlation (r)	P-value
Overall Engagement	0.580	< 0.001**
Behavioral Engagement	0.552	< 0.001**
Emotional Engagement	0.497	< 0.001**
Cognitive Engagement	0.421	< 0.001**
Pre-Test Proficiency Score	-0.312	0.001**

show greater improvement in English proficiency over the two years. Among the engagement dimensions, Behavioral Engagement emerged as the strongest predictor (r = .552), followed by Emotional Engagement (r = .497). At the same time, Cognitive Engagement showed a comparatively weaker, though still significant, relationship with proficiency gain (r = .421). This pattern suggests that students' active participation and positive affect toward the app were more strongly associated with measurable improvement than deeper cognitive processes such as strategic reflection or metacognitive regulation. In addition, the negative correlation between Pre-Test Proficiency and Proficiency Gain (r = ".312, p = .001) indicates that students with lower initial proficiency tended to experience larger gains, suggesting that the app may be particularly beneficial for mid- to lower-level learners. The strong intercorrelations among engagement dimensions (ranging from .69 to .88) further imply that behavioral, emotional, and cognitive engagement function as an interconnected system rather than isolated constructs.

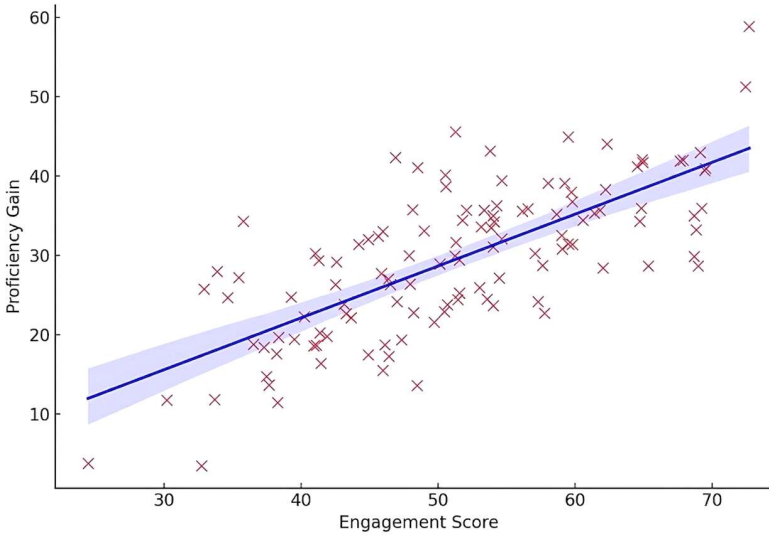

Figure 2. Pearson correlation scatterplot between engagement scores and proficiency gain

Figure 2 shows a scatterplot containing the Pearson correlation between Engagement Scores and Proficiency Gain. The red "×" sign indicates that one individual participated in the survey. The regression line in blue shows the expected direction of the scatterplot and, in this case, indicates a moderate-to-strong positive correlation. Furthermore, these students demonstrated a positive increase in proficiency gain as they engaged in more exercises with the Education Application (the app). The scatterplot also shows that students who used the app more and showed greater improvement in their English language skills were the most engaged. The positive linkage also confirms the correlation ($r = .580$, $p < .001$). Engagement in a mobile-assisted learning program is a strong predictor of language growth.

To determine the unique contribution of each engagement dimension to proficiency gain, a multiple regression analysis was conducted with Proficiency Gain as the dependent variable and Pre-Test Proficiency, Behavioral Engagement, Emotional Engagement, and Cognitive Engagement as the predictor variables. All predictors were entered simultaneously using the Enter method. The four-predictor model significantly predicted proficiency gain, $F(4, 115) = 28.42$, $p < .001$, $R^2 = .497$, Adjusted $R^2 = .481$. This indicates that approximately 49.7% of the variance in proficiency gain was explained by baseline proficiency and the three engagement dimensions. The standard error of the estimate was 3.28, indicating that predicted proficiency gain scores differed from observed scores by approximately ±3.3 points on average. Regression coefficients and model diagnostics are presented in Table 12.

The coefficient table shows that Behavioral Engagement was the strongest predictor of proficiency gain ($\beta = .39$, $p < .001$), followed by Pre-Test Proficiency ($\beta = .24$, $p = .005$) and Emotional Engagement ($\beta = .23$, $p = .013$). Cognitive Engagement was not a statistically significant predictor ($p = .275$). These results indicate that, after controlling for other variables in the model, students' behavioral and emotional engagement contributed more strongly to improvements in proficiency than cognitive engagement.

Table 12. Multiple regression predicting proficiency gain (N = 120)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.45	2.18	—	1.12	.264
	Pre-Test Proficiency	0.14	0.05	.24	2.87	.005
	Behavioral Engagement	4.12	0.98	.39	4.20	.000
	Emotional Engagement	2.19	0.87	.23	2.52	.013
	Cognitive Engagement	0.78	0.71	.09	1.10	.275

Pre-Test Proficiency also emerged as a significant predictor in the regression model ($B = 0.14$, $\beta = .24$, $p = .005$). Although the zero-order correlation in Table 12 showed a negative association between pre-test proficiency and proficiency gain, the positive regression coefficient indicates that, after controlling for the three engagement dimensions, students with higher baseline proficiency tended to make larger absolute gains. This pattern may reflect a control or suppression effect in the multivariable model and suggests that baseline proficiency should be interpreted in relation to the other predictors rather than in isolation.

Overall, the regression findings refine the earlier correlation analysis by showing that not

all engagement dimensions contributed equally to improvements in proficiency. Behavioral Engagement remained the most influential predictor, followed by Emotional Engagement, whereas Cognitive Engagement was no longer significant after accounting for the shared variance among the predictors. These findings suggest that active participation, persistence, and positive affect toward the EnglishScore app may play a more direct role in proficiency development than cognitive engagement alone.

Research Question 3: What are the perceptions of EFL learners regarding the institutional use of the EnglishScore app?

This consisted of semi-structured interviews of 10 EFL students and 5 instructors, along with 2 focus group discussions with selected

Table 13. Thematic analysis result: englishscore influenced engagement

Theme	Subtheme	Description	Sample Quotes
Theme 1: Heightened Self-Motivation and Responsibility	1.1 Score Awareness	Learners frequently checked their progress on the app, which heightened their awareness of their language progress.	*"I always checked the score. I didn't want it to decrease, and therefore, I studied more."* (Student 4)
	1.2 Ownership of Learning	Learners had a greater sense of autonomy over their own English learning, becoming self-driven and self-sufficient.	*"I felt like learning English was in my control."* (Student 9)
Theme 2: Greater Participation in the Learning Process	2.1 Confidence Development	Using EnglishScore enabled students to pinpoint their strengths and weaknesses, which in turn enabled them to participate more confidently and freely in classroom activities.	*"They understood their strengths and weaknesses. That made them more eager to take part."* (Instructor 1)
	2.2 Task Congruence with Classroom Activities	Teachers noted that students became more participative when classroom tasks were integrated with their EnglishScore outcomes.	"When the class activities matched their EnglishScore results, they were more active and willing to participate." (Instructor 4)
Theme 3: Pleasure and Game-Like experience	3.1 Gamification Effect	The app's adaptive nature and feedback in the form of points made the experience feel like a game and, therefore, was more motivating to complete.	*"It was a game, but a test. Each time, I wanted to do better."* (Student 6)
	students' self-improvement and the competitiveness of the features were reasons	The students' self-improvement and the competitiveness of the features were the reasons for continued interest over a long period.	*("I didn't get bored because I kept trying to beat my last score." (Student 7).*

participants. The interviews focused on the students' experiences with the EnglishScore app, including its usability, relevance to the course, motivational impact, and overall alignment with the course's goals. The data were transcribed and, following the procedures of thematic analysis, were grouped into four overarching themes: (1) perceived usefulness, (2) ease of use and accessibility, (3) motivational value, and (4) concerns and limitations.

Table 14. Thematic analysis – EFL learners' perceptions of the englishscore App

Theme	Subtheme	Description	Sample Quotes
1. Perceived Usefulness	1.1 Progress Tracking	Learners liked being able to monitor their language progression through CEFR-linked ratings and feedback.	*"I liked checking how my level changed each semester. It helped me know what to work on."* (Student 4)
2. Ease of Use and Accessibility	2.1 User-friendly interface	Students appreciated the app's straightforwardness and the fact that it required little, if any, instructional support to navigate.	*"It wasn't confusing like other apps. Just log in and take the test."* (Student 6)
	2.2 Mobile Accessibility	The flexibility and convenience of mobile-based design permitted users to complete assessments whenever and wherever.	*"Sometimes I used it on my phone while commuting. It fit into my schedule."* (Student 7)
3. Motivation and Engagement	3.1 Progress Motivation	Students demonstrated motivation by their ability to see their progress and compare scores.	*"I was happy when my level moved from B1 to B2. I felt proud."* (Student 8)
	3.2 Gamified Engagement	The testing felt like a game to users and motivated them to re-engage with the app to improve their scores.	*"I wanted to beat my last score, like a game."* (Student 5)
4. Concerns and Limitations	4.1 Limited Skill Coverage	Learners pointed out that the app evaluated only the receptive domains and neglected the other two: speaking and writing.	*"I wanted to practice speaking, too, though. It only tests passive skills like listening and reading."* (Student 10)

Table 15. Thematic analysis of efl educators' view of the englishscore App

Theme	Subtheme	Description	Sample Quotes
1. The Utility of the Application	1.2 The Support of Instruction	The educators used the app's diagnostic results to implement differentiated instruction and facilitate focused learning activities.	*"I used the app results to group students for activities based on their weak areas."* (Instructor 1)
	1.1 The Tracking of Proficiency	The app snapshot of learners' current proficiency levels and its reliability aided the assessment.	*"It gives me real evidence of their levels, not just my impression."* (Instructor 2)

2. Hurdles and Shortcomings	4.2 The issues of Technology and their Integration	The instructors pointed out technical issues (e.g., internet issues) and the absence of integration of classroom syllables.	*"We had Wi-Fi problems during testing, which interrupted students."* (Instructor 5)

The qualitative findings indicate a generally positive perception of the EnglishScore application among EFL learners and instructors, particularly regarding its reliability for progress monitoring and ease of use. The app promoted progress feedback and self-regulation for learners, and its gamified features appeared to enhance motivation. Concerns regarding border coverage of capabilities and technical obstacles manifest deficiencies that the app's implementation involves more formal learning settings.

Results from the study indicate that EFL learners and instructors have positive perceptions about the institutional use of the EnglishScore app. Analyzed qualitative data revealed four themes that are consistent with and build upon themes in the MALL literature: perceived usefulness, ease of use and accessibility, promotion of motivation and engagement, and worries and constraints.

Students and teachers appreciated the app's ability to track one's proficiency over time, as previously documented in the literature, and the authors have promoted the cultivation of digital proficiency. Macías Borrego (2023) and Morovvati (2024) discuss the idea that digital assessments that are diagnostic in nature can shape effective competencies and knowledge. EnglishScore also helped instructors in this study to map the learners' weaknesses, similar to the results of Schmidgall et al. (2021) who documented that app-based proficiency tests can inform data-centered teaching practices.

Contemporary research emphasizes that diagnostic assessment tools serve as powerful mechanisms for learner self-awareness and metacognitive development (Abdulaal et al., 2022). EnglishScore's ability to provide detailed skill-level feedback across grammar, vocabulary, reading, and listening domains aligns with best practices in formative assessment IRA et al. (2024)), enabling both instructors and learners to identify specific areas requiring targeted intervention. This capability is particularly valuable in institutional contexts where large-scale proficiency monitoring is essential (Hadi et al. (2023).

Accessibility facilitated the integration of EnglishScore, as it was easily accessible on mobile and had a friendly interface. This was confirmed by Qashou (2021), Al-Rahmi et al. (2021), and Al-Bashayreh et al. (2022), who identified ease of use and portability of mobile applications as primary correlates affecting learner disengagement. Recent systematic reviews underscore that mobile accessibility is a critical success factor in technology adoption, particularly in resource-constrained environments (Rezaee et al., 2020). The learners' ability to use the application on their own time is critically important in institutional contexts where students have minimal in-class exposure or face technological constraints (Rezaee et al., 2019). Furthermore, the accessibility benefits extend to supporting learner autonomy and self-regulated learning. Loewen et al. (2019) found that mobile app accessibility directly correlates with increased learner agency, particularly when learners can engage with learning tools flexibly and independently. This flexibility aligns with Self-Determination Theory principles, enabling learners to satisfy their autonomy needs (Tseng et al., 2020), a mechanism that partially explains the positive engagement outcomes observed in this study. The learners were able to use the application on their own time, which is critically important in institutional contexts where learners

have minimal exposure to the classroom or face a wide-ranging disparity in access to constraining technologies.

The positive, inspirational impact on learners that was recorded can be attributed to the app's gamified design, which also tracked learners' progress. Alzubi (2021) and Nikou & Economides (2021) stated that the mobile technologies, through immediate actionable feedback, and the design that encouraged participation, confirmed that the apps motivated learners and added to the body of knowledge. Contemporary gamification research demonstrates that progress tracking, achievement badges, and performance visualization increase intrinsic motivation in language learning contexts (Yeþilçýnar, 2023)). These design features address basic psychological needs outlined in Self-Determination Theory, particularly competence (through visible progress) and autonomy (through self-paced engagement). This also added to Ginting et al. (2024) and, Panmei & Waluyo (2022) who observed that mobile testing platforms enhanced learners' ability to own their learning and to persist in language practice.

A notable difference appeared between the quantitative and qualitative findings on cognitive engagement. In the quantitative results, Behavioral Engagement was the strongest predictor of proficiency gain, followed by Emotional Engagement. In contrast, Cognitive Engagement showed the weakest effect and was not significant in the regression model. However, the qualitative findings told a somewhat different story. Many students described experiences related to cognitive engagement, such as taking ownership of their learning, monitoring their progress, reflecting on mistakes, and feeling more self-motivated. This suggests that students were engaging cognitively with their learning, even though this was not reflected as strongly in the statistical model.

One possible explanation is that EnglishScore mainly supports practice, feedback, and progress tracking, which more directly strengthen behavioral and emotional engagement than deeper strategic thinking. In addition, cognitive engagement may be harder to capture through a questionnaire because it involves internal processes such as reflection and self-regulation, which are often expressed more clearly in interviews than in scale scores. It is also possible that the effect of cognitive engagement overlaps with that of behavioral and emotional engagement, making its unique contribution less visible in regression analysis. These findings suggest that EnglishScore is especially effective in promoting participation, persistence, and confidence. However, it may need to be combined with classroom activities such as guided reflection, goal-setting, and error analysis to strengthen deeper cognitive engagement.

App users, both learners and instructors, provided important feedback. Most important of this feedback, as cited by users, was the app's skill gaps, especially in speaking and writing. This concern is similar to the findings of Istek & Ironsi (2024), who questioned the viability of mobile proficiency apps targeting only receptive skills. Other feedback from instructors included technical difficulties and challenges of aligning the app to the curricula. Such challenges are common in studies of technology integration within institutions (Akram et al., 2022). Additionally, the social and emotional dimensions of engagement deserve emphasis. While EnglishScore is primarily an individual assessment tool, qualitative findings suggest that learners experienced a sense of achievement as their scores improved, reflecting dimensions of emotional engagement. Contemporary research on digital assessment tools indicates that when assessment results are discussed collaboratively in classroom contexts, learners experience increased relatedness and connection to peers' learning (Kaya-Capocci et

al., 2022), further enhancing overall engagement patterns.

Based on the feedback, the app's ability to provide assessment and engagement capabilities suggests that its value in EFL settings is contingent on both the app's value and the elimination of gaps in content integration and technical fit. Instructors also identified technical difficulties and challenges related to curriculum alignment. These findings are consistent with broader technology integration literature, which documents that institutional adoption of educational technologies frequently encounters implementation barriers. Carnell & Fung (2017) identified curriculum misalignment, technical support deficiencies, and inadequate teacher training as common challenges in technology integration. More recently, Gilakjani (2017) argued that successful technology integration requires not only robust technological infrastructure but also institutional commitment to professional development and curriculum redesign. The technical and curricular challenges identified by instructors in this study highlight the need for systematic institutional support beyond tool adoption.

The feedback reflects the multifaceted value of mobile assessment instruments in EFL learning environments. The findings suggest that mobile assessment instruments add significant value to instructional processes. Recent frameworks for technology integration emphasize that successful adoption requires alignment between tool affordances, institutional infrastructure, and pedagogical approaches (Abuhassna et al., 2024)).

Research by Csapó & Molnár (2019) confirms that diagnostic assessment data, when strategically leveraged, support personalized instruction and adaptive teaching practices. The positive association between EnglishScore use and learner engagement suggests that when diagnostic tools are integrated thoughtfully into curricula and supported by institutional infrastructure, they can meaningfully contribute to language learning outcomes. However, realizing this potential requires addressing both technical limitations (e.g., speaking and writing assessment, platform stability) and institutional considerations (e.g., curriculum alignment, teacher preparation).

Hybrid assessment approaches combining mobile tools with traditional assessments provide more comprehensive language proficiency evaluation and may address the gaps identified in this study. Future research should explore how institutions can strategically leverage diagnostic assessment data to personalize instruction and how mixed-modality assessment systems can better serve diverse learner needs in EFL contexts.

■ CONCLUSION

This study examined the institutional implementation of the EnglishScore mobile assessment application over two years, focusing on its association with EFL learners' proficiency development, engagement levels, and the perceptions of both students and instructors regarding its academic use. Employing a mixed-methods approach, the research aimed to provide a comprehensive understanding of the pedagogical potential and limitations of mobile-based English assessment in higher education.

The quantitative findings indicated statistically significant improvements in learners' English proficiency across grammar, vocabulary, reading, and listening. Paired-samples t-tests confirmed meaningful differences between pre-test and post-test scores, with moderate to large effect sizes observed. In addition, engagement levels were generally high, and positive correlations were identified between engagement and proficiency gain. These findings suggest that sustained engagement with the application was associated with measurable improvement in language proficiency. However, given the one-group pretest–posttest design, the results should

be interpreted as indicating association rather than direct causation.

The qualitative findings complemented the statistical results. Learners reported heightened motivation, increased responsibility for their learning, and appreciation for the app's accessibility and gamified features. Instructors valued the diagnostic function of EnglishScore, particularly its ability to provide data-informed insights into students' strengths and weaknesses. Nevertheless, participants also identified limitations, including the absence of speaking and writing assessment components, occasional internet connectivity challenges, and limited alignment with broader curricular objectives.

This research concludes that sustained use of the EnglishScore app is associated with improvements in language proficiency and higher levels of engagement among EFL learners within the institutional context. The findings suggest that the application may support language development, particularly through its user-friendly interface, structured feedback mechanisms, and motivational features. However, given the absence of a control group, the study does not establish causal effectiveness. Rather, the observed improvements occurred concurrently with the app's institutional use. Furthermore, the successful integration of such digital assessment tools into the curriculum requires careful consideration of productive language skills, particularly speaking and writing, as well as reliable technological infrastructure. Strengthening alignment between app-based assessment and broader curricular objectives may enhance its potential contribution to language learning outcomes.

The findings demonstrate a positive association between sustained use of the EnglishScore app, learner engagement, and observed proficiency gains over the two years. However, due to the absence of a control group and potential confounding variables, causality cannot be inferred. The observed gains may reflect contributions from multiple sources: app use, concurrent English instruction, academic maturation, self-selection effects, and environmental learning opportunities. Despite this limitation, the study provides valuable institutional-level evidence on technology adoption patterns and engagement mechanisms in EFL contexts. The findings support continued investment in mobile assessment tools while suggesting the need for more rigorous experimental or quasi-experimental designs in future research to establish causal effects.

DECLARATION OF GENERATIVE AI USAGE IN THE WRITING PROCESS

During the writing of this manuscript entitled "Two-Year Institutional Adoption of the EnglishScore App and Its Effects on EFL Proficiency and Engagement," the authors used ChatGPT (OpenAI) to assist with language refinement, clarity improvement, structural organization, and readability enhancement, including support in revising the Method, Literature Review, Results, and Discussion sections. All content generated with the assistance of this tool was carefully reviewed, edited, and verified by the authors to ensure that it accurately reflected their own analysis, interpretation, and conclusions. The authors assume full responsibility for the originality, integrity, and final content of the published manuscript.

REFERENCES

Abdulaal, M. A. A.-D., Alenazi, M. H., Tajuddin, A. J. A., & Hamidi, B. (2022). Dynamic vs. diagnostic assessment: impacts on EFL learners' speaking fluency and accuracy, learning anxiety, and cognitive load. *Language Testing in Asia*, *12*(1), 32.

Abuhassna, H., Alnawajha, S., Awae, F., Adnan, M., & Edwards, B. I. (2024). Synthesizing technology integration within the Addie model for instructional design: A

comprehensive systematic literature review. *Journal of Autonomous Intelligence*, *7*(5), 1–28.

Akram, H., Abdelrady, A. H., Al-Adwan, A. S., & Ramzan, M. (2022). Teachers' perceptions of technology integration in teaching-learning practices: A systematic review. *Frontiers in Psychology*, *13*, 920317.

Alam, A., & Mohanty, A. (2024). Framework of Self-Regulated cognitive engagement (FSRCE) for sustainable pedagogy: a model that integrates SRL and cognitive engagement for holistic development of students. *Cogent Education*, *11*(1), 2363157.

Al-Bashayreh, M., Almajali, D., Altamimi, A., Masa'deh, R., & Al-Okaily, M. (2022). An empirical investigation of reasons influencing student acceptance and rejection of mobile learning apps usage. *Sustainability*, *14*(7), 4325.

Ali, M. M., Mahmood, M. A., Anwar, M. N., Khan, L. A., & Hussain, A. (2019). Pakistani learners' perceptions regarding mobile assisted language learning in ESL classroom. *International Journal of English Linguistics*, *9*(4), 386–398.

Al-Rahmi, A. M., Al-Rahmi, W. M., Alturki, U., Aldraiweesh, A., Almutairy, S., & Al-Adwan, A. S. (2021). Exploring the factors affecting mobile learning for sustainability in higher education. *Sustainability*, *13*(14), 7893.

Alzubi, A. A. F. (2021). The role of mobile technologies in impacting learner autonomy in an EFL context: A systematic review. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, *11*(3), 56–73.

Anton, O. (2021). Predicting English language learners' proficiency level using EnglishScore Android Application. *Journal of Digital Ecosystem for Natural Sustainability*, *1*(2), 106–110.

Apple, M. T., Falout, J., & Hill, G. (2020). The relationship between future career self images and English achievement test scores of Japanese STEM students. *IEEE Transactions on Professional Communication*, *63*(4), 372–385.

Autin, K. L., Herdt, M. E., Garcia, R. G., & Ezema, G. N. (2022). Basic psychological need satisfaction, autonomous motivation, and meaningful work: A self-determination theory perspective. *Journal of Career Assessment*, *30*(1), 78–93.

Birajdar, L. M. (2020). Application of CEFR (Common European Framework of Reference) in Testing the English Language Proficiency Level of Polytechnic Students. *Journal of English Language Teaching*, *62*(1), 7–11.

Bouwer, R., & Dirkx, K. (2023). The eye-mind of processing written feedback: Unraveling how students read and use feedback for revision. *Learning and Instruction*, *85*, 101745.

Burston, J., & Giannakou, K. (2022). MALL language learning outcomes: A comprehensive meta-analysis 1994–2019. *ReCALL*, *34*(2), 147–168.

Carnell, B., & Fung, D. (2017). *Developing the higher education curriculum: Research-based education in practice*. UCL Press.

Champ, R. E., Adamou, M., & Tolchard, B. (2023). Seeking connection, autonomy, and emotional feedback: A self-determination theory of self-regulation in attention-deficit hyperactivity disorder. *Psychological Review*, *130*(3), 569.

Chen, J. C., Dobinson, T., & Kent, S. (2020). Students' Perspectives on the Impact of Blackboard Collaborate on Open University Australia (OUA) Online Learning. *Journal of Educators Online*,

17(1), n1.

Chiu, T. K. F. (2022). School learning support for teacher technology integration from a self-determination theory perspective. *Educational Technology Research and Development*, *70*(3), 931–949.

Csapó, B., & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in Psychology*, *10*, 1522.

Erzad, A. M. (2024). EnglishScore app practices: fostering EFL students' listening skills through English test. *Erudita: Journal of English Language Teaching*, *4*(1), 68–77.

Gilakjani, A. P. (2017). A review of the literature on the integration of technology into the learning and teaching of English language skills. *International Journal of English Linguistics*, *7*(5), 95–106.

Ginting, D., Sabudu, D., Barella, Y., Madkur, A., Woods, R., & Sari, M. K. (2024). Student-centered learning in the digital age: In-class adaptive instruction and best practices. *International Journal of Evaluation and Research in Education (IJERE)*, *13*(3), 2006–2019.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*.

Hidayati, T., & Husna, F. (2020). Learning English from Home during the COVID-19: Investigating Learners' Experience for Online and Autonomous Learning. *Langkawi: Journal of The Association for Arabic and English*, *6*(2), 202–217.

IRA, I., AGNES, F. U., & SOPIAN, S. (2024). Scoring English proficiency through Englishscore. *International Journal of Education*, *3*(1), 123–129.

Istek, E., & Ironsi, C. S. (2024). Incorporating mobile learning and the CAP (E) framework in improving the receptive skills of students. *Library Hi Tech*, *42*(2), 711–729.

Jaelani, A., Rafli, Z., & Murtadho, F. (2025). Parental Strategies in Managing L1 Interference in Early Second Language Acquisition: An Intrinsic Case Study. *JOLLT Journal of Languages and Language Teaching*, *13*(2), 616–631.

Karatas, F., & Fer, S. (2025). Breaking online language learning barriers: the effects of Experience-Based Cyclical E-Curriculum Design on learners' vocational second language performance and learning experiences: F. Karatas and S. Fer. *Educational Technology Research and Development*, 1–27.

Kaya-Capocci, S., O'Leary, M., & Costello, E. (2022). Towards a framework to support the implementation of digital formative assessment in higher education. *Education Sciences*, *12*(11), 823.

Kul, Ý. (2023). The Role of Technology in Fostering Autonomous Language Learning. *Contemporary Research In Language And Linguistics (ISSN: 2980–2253)*, *1*(2).

Kurniati, D., Riyono, A., Solaeman, A., Millan, A. R., & Noordin, N. (2025). Exploring English Proficiency Levels Of Indonesian University Students Through Englishscore Based Assessment. *Indonesian EFL Journal*, *11*(3), 539–552.

Li, T., He, P., & Peng, L. (2024). Measuring high school student engagement in science learning: an adaptation and validation study. *International Journal of Science Education*, *46*(6), 524–547.

Li, Z., & Contreras, E. S. L. (2024). Task Engagement Resources: Indicators. *Task Engagement Across Disciplines*, 223–23

Liu, Y., Razman, M. R., Zakaria, S. Z. S., Lee, K. E., Khan, S. U., & Albanyan, A. (2024). Personalized context-aware systems for sustainable agriculture development using ubiquitous devices and adaptive learning. *Computers in Human Behavior*, *160*, 108375.

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, *31*(3), 293–311.

Macías Borrego, M. (2023). Towards a digital assessment: Artificial intelligence assisted error analysis in ESL. *Macías Borrego, M.(2023). Towards a Digital Assessment: Artificial Intelligence Assisted Error Analysis in ESL. Integrated Journal for Research in Arts and Humanities*, *3*(4), 76–84.

Maier, U., & Klotz, C. (2022). Personalized feedback in digital learning environments: Classification framework and literature review. *Computers and Education: Artificial Intelligence*, *3*, 100080.

Morovvati, S. (2024). Digital Formative Assessment Strategies in Blended Learning Contexts: A Qualitative Case Study. *Assessment and Practice in Educational Sciences*, *2*(2), 1–9.

Nikou, S. A., & Economides, A. A. (2021). A framework for mobile-assisted formative assessment to promote students' self-determination. *Future Internet*, *13*(5), 116.

Okafor, J. O. (2025). The Role of Digital Tools in Assessment and Their Impact on Educational Practices. *Indonesian Journal of Innovative Teaching and Learning*, *2*(1), 58–71.

Panmei, B., & Waluyo, B. (2022). The pedagogical use of gamification in English vocabulary training and learning in higher education. *Education Sciences*, *13*(1), 24.

Qashou, A. (2021). Influencing factors in M-learning adoption in higher education. *Education and Information Technologies*, *26*(2), 1755–1785.

Ramirez-Arellano, A. (2024). Effects of personality and motivation on students' academic engagement and metacognitive-cognitive strategies. *Sage Open*, *14*(4), 21582440241295844.

Rezaee, A. A., Alavi, S. M., & Razzaghifard, P. (2019). The impact of mobile-based dynamic assessment on improving EFL oral accuracy. *Education and Information Technologies*, *24*(5), 3091–3105.

Rezaee, A. A., Alavi, S. M., & Razzaghifard, P. (2020). Mobile-based dynamic assessment and the development of EFL students' oral fluency. *International Journal of Mobile Learning and Organisation*, *14*(4), 511–532.

Sari, B. K. S. B. K., Fidyati, F., Razali, R., Rahmati, R., Husna, A., & Nursari, I. (2024). *Pemanfaatan aplikasi english score dalam peningkatkan nilai TOEFL mahasiswa tadris Bahasa Indonesia IAIN lhokseumawe* [Utilization of the English Score application to improve the TOEFL scores of Indonesian Language Education students at IAIN Lhokseumawe]. *Abdimas Siliwangi*, *7*(2), 405–414.

Schmidgall, J., Cid, J., Carter Grissom, E., & Li, L. (2021). Making the Case for the Quality and Use of a New Language Proficiency Assessment: Validity Argument for the Redesigned TOEIC Bridge® Tests. *ETS Research Report Series*, *2021*(1), 1–22.

Shadiev, R., Hwang, W.-Y., & Huang, Y.-M. (2017). Review of research on mobile language learning in authentic environments. *Computer Assisted Language Learning*, *30*(3–4), 284–303.

She, L., Khoshnavay Fomani, F., Marôco, J., Allen, K.-A., Sharif Nia, H., & Rahmatpour, P. (2023). Psychometric properties of the university student engagement inventory among Chinese students. *Asian Association of Open Universities Journal*, *18*(1), 46–60.

Shen, H., Ye, X., Zhang, J., & Huang, D. (2024). Investigating the role of perceived emotional support in predicting learners' well-being and engagement mediated by motivation from a self-determination theory framework. *Learning and Motivation*, *86*, 101968.

Tran, T. T. T., Nguyen, P. L. A., Nguyen, N. H. T., & Tran, D. T. (2021). An Empirical Study on Students' Perception and Satisfaction Towards Online Assessment and Testing in Tertiary Education. *18th International Conference of the Asia Association of Computer-Assisted Language Learning (AsiaCALL–2-2021)*, 207–216.

Tseng, W.-T., Liou, H.-J., & Chu, H.-C. (2020). Vocabulary learning in virtual environments: Learner autonomy and collaboration. *System*, *88*, 102190.

Utami, B. R., & Wahyudin, A. Y. (2022). Does self-esteem influence student english proficiency test scores? *Journal of English Language Teaching and Learning*, *3*(2), 16–20.

Wong, R. (2020). When no one can go to school: does online learning meet students' basic learning needs? *Interactive Learning Environments*, 1–17.

Yeþilçýnar, S. (2023). Personalized Learning Through Gamification: A ChatGPT Approach to English Language Learning. In *Transforming the Language Teaching Experience in the Age of AI* (pp. 44–64). IGI Global.

Zare, J., Delavar, K. A., Derakhshan, A., & Pawlak, M. (2024). The relationship between self regulated learning strategy use and task engagement. *International Journal of Applied Linguistics*, *34*(3), 842–861.

Zhang, Z. (2022). Promoting student engagement with feedback: insights from collaborative pedagogy and teacher feedback. *Assessment & Evaluation in Higher Education*, *47*(4), 540–555.

Zolfaghari, Z., Karimian, Z., Zarifsanaiey, N., & Farahmandi, A. Y. (2025). A scoping review of gamified applications in English language teaching: a comparative discussion with medical education. *BMC Medical Education*, *25*(1), 274.