Journal of Sustainable Development and Policy

Volume: *3;* ***Issue:*** *4*
Pages: *01– 31*

AI-BASED LANGUAGE EDUCATION PLATFORMS: A SYSTEMATIC ANALYSIS OF EDTECH TOOLS FOR ENGLISH PROFICIENCY

Elmoon Akhter[1];

[1]. MA in English, State University of Bangladesh, Dhaka, Bangladesh.
Email: elmoonshimu@gmail.com

Doi: 10.63125/5cdy4608
Received: *29 September 2024;* ***Revised:*** *20 October 2024****; Accepted:*** *18 November 2024****; Published:*** *24 December 2024*

Abstract

This study addresses a practical problem in higher education and enterprise deployments: institutions are adopting AI-based language learning platforms without clear, quantified evidence of which features most strongly relate to measurable English proficiency. The purpose is to estimate feature–outcome relationships in authentic settings. Using a quantitative, cross-sectional, case-based design, we analyze multi-institution data from cloud-hosted, enterprise-grade platforms (writing focused, speaking focused, and integrated skills). The sample comprises linked survey, telemetry, and assessment records, with proficiency indicators mapped to CEFR bands or rubric scores. Key variables include engagement intensity, feedback specificity, feedback immediacy, adaptivity breadth, and spacing quality, with controls for baseline proficiency, prior exposure, study time, device access, and demographics. The analysis plan includes descriptive statistics, zero-order correlations, multiple linear regression with case fixed effects and heteroskedasticity-robust errors, moderation by motivation, subgroup contrasts by platform modality, and spline checks for nonlinearity. Headline findings indicate that feedback specificity shows the largest unique association with proficiency (about 0.17 SD per 1 SD increase), followed by engagement (≈0.13), spacing quality (≈0.10), feedback immediacy (≈0.09), and adaptivity breadth (≈0.08). Motivation positively moderates the engagement–proficiency link, speaking cases benefit most from fast feedback and spaced practice, and writing cases benefit most from granular, itemized feedback. Returns to weekly minutes flatten beyond the upper quintile, suggesting calibrated rather than maximal practice. Implications for policy and practice include engineering actionable feedback at scale, setting latency service targets, exposing spacing indices, and aligning dashboards with revision uptake and error convergence to guide instruction and procurement.

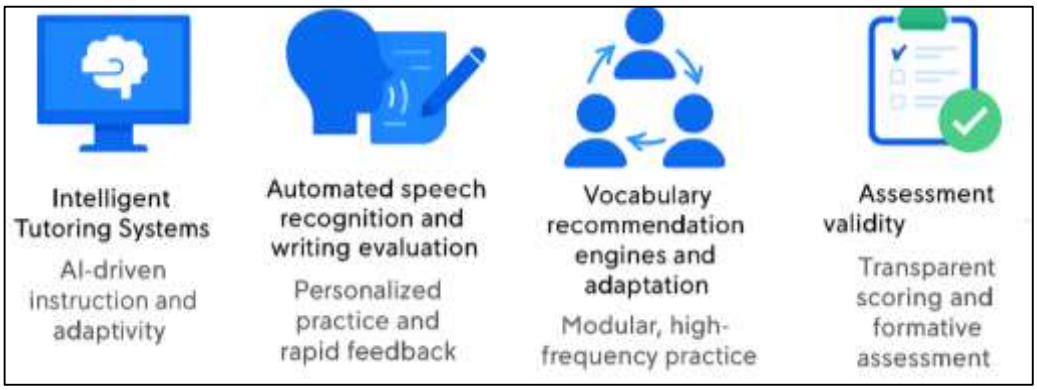
Keywords

AI-Based Language Learning; Feedback Specificity; Learner Engagement; Spacing Effect; English Proficiency;

INTRODUCTION

Artificial intelligence (AI) in language education broadly refers to computational methods that learn from data to support or automate instructional tasks such as tutoring, feedback, assessment, and personalization within computer-mediated environments (Fleckenstein et al., 2023). In this study, AI-based language education platforms are digital systems that embed intelligent tutoring, natural-language processing (NLP), automated speech recognition (ASR), automated writing evaluation (AWE), recommendation, and/or adaptive assessment to promote English proficiency across reading, writing, listening, speaking, vocabulary, and integrated skills. The international significance of English proficiency is well documented in higher education and the global knowledge economy: the expansion of English-medium instruction (EMI) in universities has become a structural feature of internationalization, with documented growth across regions and disciplines (Kulik & Fletcher, 2016).

Figure 1: AI-Based Language Education Platforms



Concurrently, ministries and institutions seek cost-effective, scalable technologies to meet rising demand for high-quality, data-driven language learning needs that align squarely with AI's comparative advantages in adaptivity, feedback timeliness, and real-time analytics (Derwing & Munro, 2005). From a theoretical perspective, the CALL–SLA interface positions AI platforms as mediators of input, interaction, output, and feedback processes that underpin acquisition, provided their tasks are appropriately aligned to learning goals (Chapelle, 2009). Against this backdrop, the proposed quantitative, cross-sectional, case-study–based design examines AI platforms' contribution to English proficiency via descriptive statistics, correlation, and regression modeling grounded in Likert-type measures. By consolidating multi-country evidence on AI tutoring (e.g., AutoTutor), automated feedback (AWE), and ASR-mediated speaking practice, this introduction establishes conceptual and empirical foundations for testing hypothesized relationships between platform features, learner engagement, and proficiency outcomes in real institutional settings (D'Mello & Graesser, 2015a).

Historically, intelligent tutoring systems (ITS) pioneered the integration of AI into instruction by modeling domain knowledge, learner states, and pedagogical strategies to deliver adaptive guidance. Meta-analytic evidence indicates that ITS produce learning gains of roughly two-thirds of a standard deviation over conventional instruction underscoring the potential relevance of AI-driven adaptivity within language contexts (Ngo et al., 2023). In language education specifically, dialogue-based tutors such as AutoTutor illustrate how mixed-initiative tutorial dialogue, deep reasoning prompts, and feedback mechanisms can be aligned with construct-relevant tasks and measurable learning outcomes (Graesser et al., 2005). The broader technology-supported language learning literature further finds that computer-supported pedagogies, when compared to non-technology conditions, yield small but significant positive effects across decades of research syntheses (Grgurović et al., 2013). These converging findings situate modern AI platforms within a mature lineage of evidence-based instructional technologies whose effects depend on design quality (task authenticity, alignment, feedback specificity) and implementation fidelity in classrooms. In English proficiency initiatives especially those scaling EMI AI platforms provide individualized practice and rapid feedback that are otherwise costly to deliver at scale. For program designers and policy makers, the implication is that

platform choice should be empirically grounded: selection and implementation plans ought to consider the nature of the AI intervention (tutoring vs. feedback vs. recognition), the targeted proficiency construct, and the evidence base supporting each feature for the learner population under study. The present research operationalizes these considerations into testable models relating platform features and usage to proficiency indicators (Hwang et al., 2020).

Two AI modalities have become especially salient for English proficiency: automated writing evaluation (AWE) and automatic speech recognition (ASR). AWE uses NLP and machine-learned scoring models to provide holistic/analytic scores and formative feedback on grammar, cohesion, organization, and mechanics. Earlier classroom studies and reviews documented both opportunities and challenges of AWE use in writing instruction, emphasizing the need for pedagogically mediated deployment (Warschauer & Grimes, 2008). More recently, a large-scale meta-analysis reported positive, significant effects of automated feedback on L2 writing quality, with effects moderated by feedback scope and integration into instruction (Fleckenstein et al., 2023). Complementarily, ASR-based pronunciation and speaking tools leverage speech technologies to offer immediate, individualized feedback at scale. A 2023 meta-analysis in ReCALL synthesized 15 studies (38 effect sizes) and found a medium overall effect ($g \approx 0.69$) of ASR on ESL/EFL pronunciation performance; explicit corrective feedback and peer-mediated practice were associated with larger gains (Ngo et al., 2023). These developments mark a methodological inflection point: AI no longer merely scores or drills but participates in adaptive formative assessment, wherein feedback timing, specificity, and difficulty progression can be tuned to learner profiles and course objectives. The present study incorporates these modalities into its construct model, enabling the statistical testing of relationships between platform usage (AWE/ASR features), perceived feedback usefulness, and standardized or curriculum-aligned proficiency indicators (OECD, 2019).

Beyond writing and speaking, AI-enabled platforms address vocabulary and integrated-skills development through recommendation engines and micro-adaptivity. Meta-analyses show that technology-assisted L2 vocabulary learning yields significant effects, with design choices (e.g., spacing, multimodality, retrieval practice) moderating outcomes (Yu & Trainin, 2021). Mobile-assisted language learning (MALL) research likewise reports moderate positive effects on language learning performance; as devices have become ubiquitous, mobile contexts have unlocked flexible, high-frequency practice that is amenable to data-driven personalization (Sung et al., 2015). At the ecosystem level, comprehensive reviews in higher education highlight that AI and educational technologies influence not only achievement but also assessment practices, learner autonomy, and institutional capacities for analytics-informed decision-making (Heffernan et al., 2014). Within English proficiency initiatives aligned to EMI and internationalization agendas, these strands converge: AI platforms can sequence evidence-based learning events (e.g., spaced vocabulary recycling, targeted reading micro-skills) and link them to feedback loops, thereby producing data suitable for correlational and regression analyses of proficiency gains. This study's survey instrumentation and platform telemetry variables are therefore conceptualized to capture both perceived and observed adaptivity and feedback quality, enabling robust tests of the proposed hypotheses concerning proficiency outcomes (Li et al., 2022).

Any examination of AI platforms for English proficiency must also reckon with assessment validity and fairness when automated scoring or recognition is involved. The e-rater® research program reported reliability/validity evidence for automated essay scoring, clarifying how limited feature sets and transparent modeling can support score interpretability (Attali & Burstein, 2006). In pedagogical contexts, however, earlier school-based AWE deployments showed that benefits often hinge on teacher mediation and integration into process writing, cautioning against uncritical automation (D'Mello & Graesser, 2015b). On the speaking side, ASR-mediated pronunciation training complements long-standing research on intelligibility and comprehensibility in second-language speech, where instruction and feedback rather than accent reduction per se are central to communicative outcomes (Sanjid & Farabe, 2021; Popenici & Kerr, 2017). Recent meta-analytic evidence specific to ASR confirms that explicit corrective feedback and collaborative practice produce stronger effects than dictation-style use, underscoring the formative assessment role of AI tools (Sun, 2023). For this study's model, we therefore treat automated scores not as high-stakes summative outcomes but as proximal indicators that mediate relationships between platform use and validated proficiency measures. This distinction

is critical for interpreting correlations and regression coefficients meaningfully within a cross-sectional design spanning diverse learners and institutional contexts (Macaro, 2018; Omar & Rashid, 2021).

The international policy environment strengthens the rationale for interrogating AI platforms' links to proficiency. EMI's rapid expansion in higher education underscores the role of English proficiency as both a gatekeeping and enabling factor for academic participation, mobility, and scholarly communication (Macaro et al., 2018; Zaman & Momena, 2021). System-level monitoring (e.g., OECD's *Education at a Glance*) documents cross-national pressures on tertiary education quality, completion, and labor-market alignment conditions that elevate demand for scalable, cost-effective language support (OECD, 2019). Within this milieu, AI-based platforms compete for adoption on promises of personalization, analytics, and efficiency. Yet systematic reviews of AI in higher education caution that deployment must be anchored in rigorous evidence, ethical data governance, and attention to pedagogical alignment (Rony, 2021; Zawacki-Richter et al., 2019). These concerns structure the present study's objectives and hypotheses: to quantify associations between platform features (e.g., tutor adaptivity, feedback immediacy), learner engagement (usage intensity, perceived usefulness), and proficiency indicators; and to estimate the unique contribution of AI features after accounting for learner covariates and contextual factors in multivariate models. By using case-study sampling across institutions that have already integrated AI platforms into routine instruction, the study connects policy-level imperatives to ground-level outcomes in a way that is statistically tractable and externally relevant. (Zaki, 2021; Zhai & Ma, 2023).

Finally, the proposed research is positioned within an evolving evidence base that spans meta-analyses of technology-assisted learning, domain-specific AI interventions, and theoretical accounts of the CALL–SLA nexus. Syntheses indicate that technology-supported conditions generally outperform or match traditional instruction, with effect sizes contingent on design and alignment (Fleckenstein et al., 2023). Within AI-specific subdomains, dialogue-based tutoring (AutoTutor), AWE for writing, and ASR for speaking show positive effects under pedagogically coherent implementation (Attali & Burstein, 2006; Hozyfa, 2022). These literatures collectively justify a modeling strategy that treats AI features as predictors and English proficiency as outcomes, while controlling for learner background and engagement variables. The seven-paragraph introduction has defined key constructs, situated the study's global relevance, and outlined how prior research informs the present hypotheses. The subsequent sections (method, measures, data analysis) will operationalize constructs into Likert-type scales and platform usage metrics suitable for descriptive statistics, correlation analyses, and regression modeling in a cross-sectional, case-study frame (Derwing & Munro, 2005; Hwang et al., 2020; Arman & Kamrul, 2022).

The objective of this study is to quantify the relationships between specific features of AI-based language education platforms and measurable indicators of English proficiency within authentic institutional contexts, using a cross-sectional, multi–case study design and a strictly quantitative analytic plan. Concretely, the study aims to (a) operationalize platform engagement through logged or self-reported usage intensity, session frequency, and task completion ratios; (b) derive feature-level indices for automated writing evaluation, automatic speech recognition, adaptive sequencing, and feedback immediacy/precision; and (c) measure learner-reported constructs motivation, self-efficacy, perceived usefulness, perceived feedback quality, and usability via Likert five-point scales with established reliability thresholds. A first objective is to describe, with summary statistics, the distributional properties of engagement and feature utilization across multiple platforms and institutions, disaggregated by baseline proficiency bands and demographic covariates, thereby providing a transparent portrait of how learners actually use AI functions in routine study. A second objective is to estimate zero-order associations among engagement, feature indices, learner-reported constructs, and proficiency outcomes (standardized scores or rubric-based speaking/writing ratings) to identify bivariate patterns that merit multivariate testing. A third objective is to specify and fit multiple regression models in which proficiency outcomes are predicted by engagement and feature indices while adjusting for baseline proficiency, prior exposure to English, study time outside the platform, device access, and demographic controls, with heteroskedasticity-robust standard errors and diagnostic checks for multicollinearity and influential observations. A fourth objective is to examine moderation by motivation and self-efficacy through interaction terms with engagement, enabling an

assessment of whether the strength of association between platform use and proficiency differs across learner profiles. A fifth objective is to conduct platform-level and subgroup analyses (e.g., beginners versus intermediate learners; writing-focused versus speaking-focused users) to assess the stability of coefficients across cases and tasks. A sixth objective is to validate the internal consistency of all multi-item scales and to document construct validity evidence through item–total correlations and factor-analytic summaries when applicable. Together, these objectives specify a coherent empirical agenda that moves from descriptive mapping to correlational screening and into controlled modeling, yielding statistically defensible estimates of the extent to which distinct AI features and engagement patterns are associated with English proficiency within the defined population and sampling frame.

LITERATURE REVIEW

The literature on AI-based language education platforms spanning intelligent tutoring systems, automated writing evaluation, automatic speech recognition, adaptive assessment, and data-driven recommendation coalesces around a core question: how do specific technological affordances align with well-established mechanisms of second language acquisition to support measurable gains in English proficiency? Foundational work in computer-assisted language learning anchors this inquiry in theories of input, interaction, output, and feedback, while more recent strands in learning analytics, mobile-assisted language learning, and human–AI collaboration emphasize personalization, immediacy of formative feedback, and scalable practice across contexts. Across decades, meta-analyses of technology-supported instruction generally report small-to-moderate positive effects, but the heterogeneity of tasks, durations, learner profiles, and implementation fidelity complicates direct attribution to any single feature. Within this mosaic, automated writing evaluation and ASR-based speaking tools have emerged as focal subfields: AWE studies interrogate feedback scope and quality (surface-level versus discourse-level guidance), student uptake, and teacher mediation, whereas ASR research concentrates on intelligibility-focused practice, corrective feedback timing, and integration into communicative activities. Parallel developments in adaptive sequencing and recommendation algorithms seek to optimize practice schedules and difficulty progression, often operationalized through mastery estimates and micro-adaptive task selection, yet questions remain about construct coverage and alignment with validated proficiency frameworks. Importantly, much of the evidence base is fragmented mixing lab studies with classroom deployments, short interventions with semester-length implementations, and proprietary analytics with standardized assessments leading to a patchwork of outcome measures and effect sizes that challenge synthesis. The emerging literature on learner psychology in AI environments adds another layer, examining motivation, self-efficacy, and perceived usefulness as mediators or moderators of performance, while usability and accessibility considerations shape real-world engagement. Taken together, these strands motivate a literature review that first maps the conceptual terrain linking AI features to SLA-relevant processes, then synthesizes empirical findings by modality and construct, and finally identifies methodological patterns measurement validity, sampling strategies, and statistical controls that condition the credibility and comparability of reported outcomes. This orientation sets the stage for a focused, quantitative investigation that treats platform engagement and feature utilization as predictors, proficiency indicators as outcomes, and learner/contextual factors as covariates within a transparent, reproducible modeling framework.

Theoretical Foundations

A theory-driven account of AI-based language education platforms begins with how instructional feedback and interaction catalyze change in learners' developing interlanguage. In classroom second-language acquisition, corrective feedback exerts statistically reliable, durable effects on target-language development, with prompts often outperforming recasts because they recruit deeper processing and self-repair; this provides a mechanism for why AI systems that deliver immediate, specific signals can shift performance distributions on grammar, lexis, and discourse measures (Lyster & Saito, 2010). In parallel, a design science of feedback clarifies that formative feedback information that is timely, task-specific, and nonjudgmental optimally supports learning when it (a) targets misconceptions, (b) indicates how to improve, and (c) is delivered at a grain size matched to the learner's current representation (Mohaiminul & Muzahidul, 2022; Shute, 2008). Together, these strands tie core AI affordances (e.g., instant error highlighting, exemplars, graduated hints, adaptive re-sequencing) to

processing operations stipulated by SLA theory: noticing, restructuring, and automatization within meaningful tasks. Under this view, "intelligent" behavior is not merely algorithmic sophistication but instructional alignment: an AI platform is most plausibly effective when its feedback forms, timing, and contingency match both the linguistic target and the learner's proximal zone of development (Omar & Ibne, 2022). Consequently, in the present study, feedback immediacy, specificity, and adaptivity are treated as feature-level constructs expected to covary with proficiency outcomes because they instantiate theoretically favored conditions for uptake. This theoretical framing also motivates measuring perceived feedback quality in tandem with observed usage: learners' judgments of usefulness can index whether feedback is experienced at the right level of diagnosticity, while platform logs indicate whether such feedback is received at the right moments in the learning sequence (Lyster & Saito, 2010; Shute, 2008).

Figure 2: Theoretical Foundations of AI-Based Language Education Platforms

A second pillar derives from the cognitive psychology of practice scheduling. The spacing effect the finding that learning distributed across time outperforms massed practice has been repeatedly confirmed and quantified, with optimal interstudy intervals scaling with the desired retention horizon (Cepeda et al., 2006; Hasan, 2022). In language learning, spacing supports durable consolidation of vocabulary, morphosyntax, and formulaic sequences by allowing partial forgetting and effortful retrieval processes that strengthen memory traces and discrimination among competitors. Modern platforms operationalize spacing via scheduler policies (e.g., expanding intervals, difficulty-weighted review), so an empirically grounded theory proposes that more calibrated spacing indexed by an alignment between interval growth and learner accuracy should predict higher proficiency (Mominul et al., 2022). A complementary evidence base on high-utility learning techniques ranks retrieval practice and distributed practice as especially potent for long-term retention and transfer, suggesting that AI systems combining frequent, low-stakes retrieval with adaptive spacing should yield stronger outcomes than systems emphasizing passive exposure (Dunlosky et al., 2013; Rabiul & Praveen, 2022). For measurement, this implies capturing not only raw activity counts but also temporal structure in engagement (e.g., median intersession interval, variance of spacing, proportion of reviews occurring near an item's predicted forgetting point). Analytically, these features can be modeled as predictors of standardized proficiency indicators or rubric-based writing/speaking scores, with the expectation that, controlling for baseline level and exposure, greater adherence to spaced retrieval regimes associates with higher performance (Farabe, 2022; Roy, 2022). Thus, the cognitive scheduling literature supplies testable, time-sensitive parameters that link platform telemetry to proficiency variance within and across cases (Dunlosky et al., 2013; Shute, 2008).

A third pillar concerns learner–task fit and cognitive load. The expertise reversal effect predicts that instructional formats advantageous for novices (e.g., highly guided, worked-example–rich sequences) can become suboptimal as knowledge grows, because redundant guidance imposes extraneous load and suppresses self-explanation (Kalyuga, 2007; Rahman & Abdul, 2022; Razia, 2022). In AI environments, this implies that static levels of scaffolding may hinder intermediate learners, whereas adaptive fading of hints and expansion of problem spaces should improve efficiency and transfer. Accordingly, we conceptualize adaptivity not only as item difficulty selection but also as guidance modulation: the platform's ability to taper feedback verbosity and shift from form-focused to integrative tasks as competence increases. To render these ideas estimable, the study treats proficiency as a continuous outcome Yi predicted by engagement and feature indices X, including a spacing-sensitive scheduler index and a guidance-modulation index, while allowing for interactions with motivation/self-efficacy. A schematic model is

$$Proficiency_i = \beta_0 + \beta_1 Engagement_i + \beta_2 AdaptiveFeedback_i + \beta_3 SpacingIndex_i + \beta_4 GuidanceModulation_i + \beta_5 (Engagement_i \times Motivation_i) + \gamma^{\top} Controls,$$

where controls include baseline proficiency, prior exposure, device access, and study time. Within this framework, significant positive β_2–β_4 would align with feedback, spacing, and expertise-reversal predictions, while a positive β_5 would indicate that motivated learners realize greater marginal returns from engagement a theoretically coherent moderation. By grounding feature selection, variable construction, and model specification in established mechanisms, the section articulates why particular AI capabilities should relate to measurable gains and how those relations can be captured in cross-sectional, multi-case regression analyses (Kalyuga, 2007; Zaki, 2022).

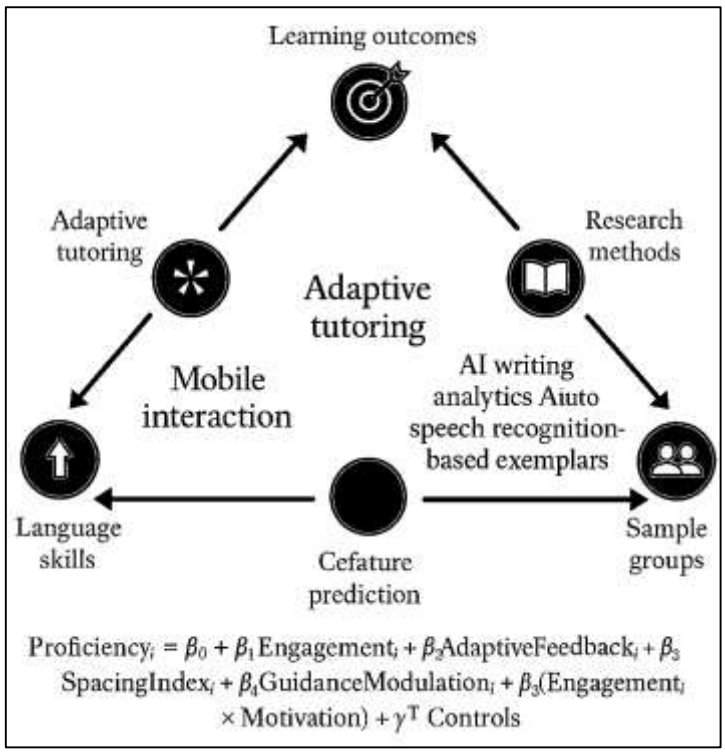
AI-Enabled Modalities and Feature Taxonomies in Language EdTech

A coherent view of AI-based language education platforms begins by mapping the feature space that most directly supports English proficiency development. Core modalities typically include intelligent tutoring (task selection, hinting, mastery estimation), natural language processing for writing support (formative feedback on accuracy, discourse, and cohesion), speech technologies for listening and speaking (recognition, scoring, and targeted recasts), and recommendation engines that orchestrate spaced review and retrieval practice. Positioning these modalities within a cumulative evidence base benefits from two anchor perspectives. First, research on tutoring systems clarifies how adaptive guidance, contingent feedback, and stepwise problem solving can approximate some functions of human tutoring in scalable ways; this work frames the kinds of feature–outcome relations one should expect when platforms personalize sequence and feedback at the level of specific subskills (Kanti & Shaikat, 2022; VanLehn, 2011). Second, methodological standards for interpreting effect sizes in second-language studies help translate platform-linked gains into practically interpretable terms, encouraging the reporting of standardized effects, confidence intervals, and model diagnostics alongside raw score changes so that feature contributions remain comparable across contexts and assessment instruments (Danish, 2023a, 2023b; Plonsky & Oswald, 2014). For the present review, this taxonomy is operationalized as measurable indices engagement intensity, feedback immediacy/specificity, adaptivity breadth, and schedule quality each hypothesized to co-vary with proficiency when learning activities align with constructs in reading, writing, listening, speaking, and vocabulary.

Mobile and connected learning environments have expanded the ecological footprint of those AI modalities. As smartphones, sensors, and ubiquitous connectivity normalize just-in-time practice, platforms can distribute micro-activities across authentic settings, embed immediate feedback into short interaction bursts, and log high-resolution traces of learner behavior. Reviews of language learning on smartphones synthesize how mobility affords frequent, brief, and situated encounters with target forms, thereby increasing opportunities for input, output, and interaction that AI services can personalize and analyze at scale; this mobile substrate also broadens participation in contexts where desktop access is limited and supports hybrid classroom–self-study workflows (Godwin-Jones, 2017; Muhammad & Redwanul, 2023). Meanwhile, a meta-analytic lens on synchronous computer-mediated interaction underscores that technology-mediated communication can support developmentally relevant interactional work negotiation of meaning, feedback uptake, and pushed output providing a theoretical bridge from classic interaction research to conversational agents and AI-augmented chat environments that script prompts, regulate turn-taking, and deliver contingent feedback (Razia, 2023;

Reduanul, 2023; Ziegler, 2016). Within this mobile-interaction nexus, the practical question for researchers becomes how to parse telemetry into constructs that are psychometrically meaningful (e.g., spacing indices, feedback latency distributions) and statistically estimable in models linking feature use to proficiency outcomes.

Figure 3: AI-Enabled Modalities and Feature Taxonomies in Language EdTech



A complementary strand anchors AI writing and vocabulary supports in data-driven learning (DDL), where learners engage directly with usage evidence collocations, patterns, and genre-specific moves via corpora and concordances (Sadia, 2023; SSrinivas & Manish, 2023). A large meta-analysis of corpus use in language learning reports that DDL can yield positive effects across skills and proficiency levels, especially when tasks scaffold search, noticing, and pattern abstraction; this offers a principled rationale for NLP-enabled platforms that surface corpus-derived exemplars, automate pattern checks, and tailor feedback to authentic usage (Boulton & Cobb, 2017; Mesbaul, 2024; Omar, 2024; Zayadul, 2023). Conceptually, these DDL-inspired capabilities complement adaptive tutoring, ASR feedback, and mobile interaction by addressing lexical and discourse development through exposure to representative input and structured discovery. For measurement, the same principles that support rigorous effect interpretation in second-language research encourage platforms and studies to report standardized proficiency outcomes and transparent mappings from usage events to learning constructs, enabling synthesis across heterogeneous tools and settings (Momena & Praveen, 2024; Muhammad, 2024; Noor et al., 2024). In sum, the literature converges on a feature taxonomy adaptive tutoring, AI writing analytics, ASR-mediated speaking practice, mobile orchestration, and DDL-based exemplars that is theoretically grounded in interaction, feedback, and usage-based learning and is practically amenable to quantitative modeling in cross-sectional, multi-case designs. (Godwin-Jones, 2017).

Measurement in Operationalizing English Proficiency

Building cumulative evidence about AI-based language education platforms requires outcome measures whose interpretations are defensible, generalizable, and comparable across contexts. Validity theory treats test scores as *inferences* about underlying constructs, supported by a network of arguments concerning content representativeness, internal structure, relations to other variables, and consequences of use; this argument-based approach is especially salient when researchers link platform features (e.g., adaptive feedback, spacing quality) to proficiency metrics harvested from classroom

assessments or embedded platform tasks (Kane, 2013). In practice, construct definitions should specify which facets of English proficiency linguistic accuracy, discourse organization, listening comprehension, or interactive speaking are being targeted and how those facets are sampled by items or tasks. Internal-structure evidence can be explored by examining whether multi-item scales (e.g., perceived feedback quality, motivation) exhibit the dimensionality and item functioning implied by the construct model. Externally, correlational patterns offer nomological checks: platform-derived indicators should relate to outcome measures in ways that align with theory (e.g., stronger associations for feedback-sensitive skills). Finally, consequences evidence matters whenever scores are used to make programmatic judgments or allocate support; even in a research design that is cross-sectional rather than experimental, the interpretive argument should recognize that choices about rubrics, raters, and automated scoring parameters can shape who appears to benefit from a given AI feature and why (Kunnan, 2018). By making the validity argument explicit, the present review positions proficiency outcomes not as fixed facts but as *claims* that require documentation and scrutiny before they enter regression models relating features and engagement to achievement.

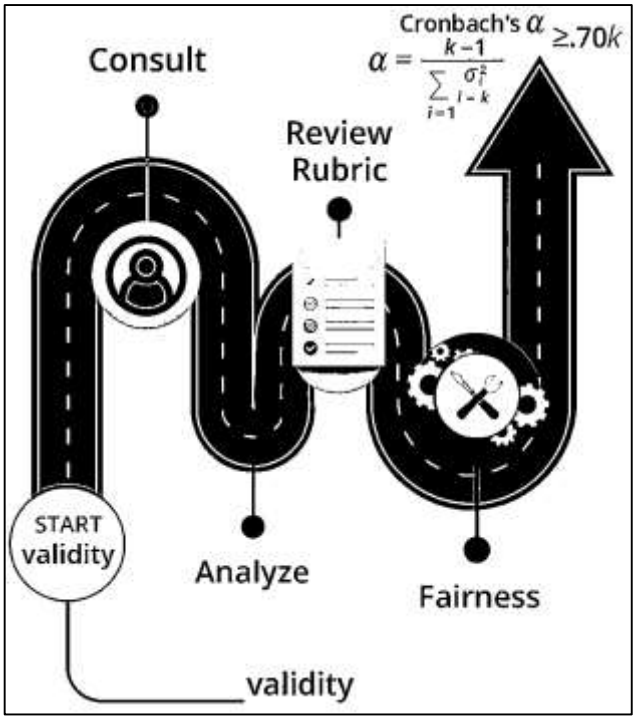
Reliability and score precision are foundational to that validity argument. Because the study employs Likert-type scales for learner-reported constructs and rubric-based ratings for writing or speaking, internal-consistency and interrater indices must be reported and interpreted with care. Cronbach's alpha remains a widely used summary of internal consistency, but its interpretation depends on assumptions (tau-equivalence, unidimensionality) that are often violated in practice; as a result, alpha can both under- and over-estimate reliability if items vary in discrimination or form multiple factors (Sijtsma, 2009). When alpha is used, it should be accompanied by transparent item analyses and, where feasible, complementary estimates. Formally, for a k-item scale with total-score variance σ_t^2 and item variances σ_i^2, the coefficient is

$$\alpha = \frac{k-1}{k}\left(1 - \frac{\sigma_T^2}{\sum_{i=1}^{k}\sigma_i^2}\right),$$

which highlights how internal consistency rises as common covariance among items grows. In applied settings, target thresholds (e.g., α≥.70\alpha \ge .70α≥.70) are rules of thumb rather than absolutes; researchers should focus on whether the reliability is sufficient for *group-level* inferences in correlational and regression analyses, and whether subscales behave as intended across proficiency bands (Tavakol & Dennick, 2011). For rater-mediated outcomes, decision quality depends on both rater agreement and scale functioning; designs should therefore document rater training, provide evidence of consistent category use, and consider many-facet or generalizability approaches when feasible. In sum, reliability reporting serves not as a mere checklist item but as a quantitative warranty that the observed relations between platform features and proficiency are not artifacts of noisy measurement (Plonsky, 2015).

Fairness provides a third, indispensable lens for interpreting proficiency outcomes in AI-enhanced settings. Fairness and justice in language assessment extend beyond the absence of bias in items to encompass the entire testing ecosystem, including accessibility, consequential validity, and the equity of decisions informed by scores (Kunnan, 2018). For research that draws inferences from platform-linked outcomes, this implies routine checks for subgroup performance differences not explained by construct-relevant variance for example, comparing residuals across demographic groups after controlling for baseline proficiency and exposure. Analytically, a straightforward approach is to include protected-group indicators and their interactions with key predictors in the regression model and to examine whether coefficients governing feature–outcome relations remain stable; significant interactions may signal differential effectiveness or potential measurement noninvariance that warrants follow-up. Transparency also extends to effect reporting: standardized effect sizes, confidence intervals, and model diagnostics help situate findings within the broader L2 methods literature, which has called for stronger quantitative rigor to improve replicability and meta-analytic integration (Plonsky, 2015). Bringing these strands together, the study treats validity (what the scores mean), reliability (how consistently they are estimated), and fairness (for whom the inferences hold) as interlocking criteria. The resulting measurement framework makes it possible to compare AI feature indices and engagement variables across cases without inflating claims, thereby supporting cautious, quantitatively grounded interpretations of how tutoring, feedback, and scheduling capabilities relate to observed proficiency profiles (Kunnan, 2018).

Figure 4: Measurement in Operationalizing English Proficiency



Learning Analytics and Telemetry for Modeling Platform–Proficiency Relations

A robust account of AI-based language education platforms requires turning raw telemetry clickstreams, session intervals, hint requests, and automated feedback events into constructs that meaningfully index engagement, adaptivity, and assessment-relevant behavior. Learning analytics provides the conceptual and methodological scaffolding for this transformation by articulating how institutional goals, measurement choices, and data flows interact in cyclical processes of collection, analysis, and interpretation. In higher education settings, early work demonstrated that even comparatively coarse learning management system traces could forecast achievement and flag at-risk learners when summarized as interpretable indicators (e.g., frequency, recency, and diversity of activity), establishing the feasibility of data-informed monitoring at scale (Macfadyen & Dawson, 2010). Complementarily, field syntheses described the emerging drivers and challenges of learning analytics, emphasizing the need to align indicators with pedagogical intent and to represent outcomes in ways that support decisions by multiple stakeholders (Ferguson, 2012). For AI-enhanced language learning, these principles map naturally onto our constructs: feedback immediacy and specificity become event-level features; adaptivity breadth becomes a distribution over difficulty transitions; spacing quality becomes a temporal pattern correlating review intervals with prior performance; and engagement acquires a time-sensitive profile rather than a single cumulative count. When the analytics pipeline is articulated in this way, it becomes possible to connect platform features to proficiency measures through defensible operationalization, transparent summary statistics, and model-ready variables that reflect substantive theory rather than arbitrary logs (Ferguson, 2012; Macfadyen & Dawson, 2010).

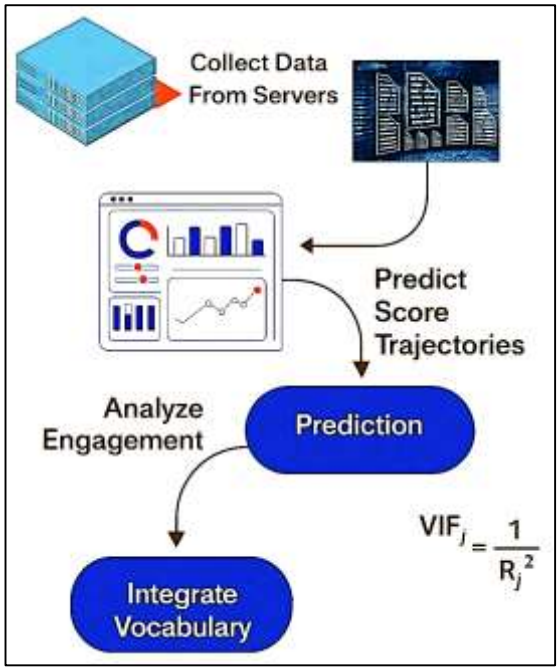
Behavioral patterns in large, digitally mediated courses offer additional leverage for constructing feature indices that travel well across platforms. Research on massive open online courses, for instance, decomposed disengagement into distinct trajectories e.g., early drop-off, intermittent participation, and consistent engagement demonstrating that temporal dynamics and sequence regularities are at least as informative as totals for predicting outcomes (Kizilcec et al., 2013). Educational data mining surveys systematized algorithmic approaches for discovering such structures (classification, clustering, sequential pattern mining, and association rules), while underscoring the importance of interpretability and domain alignment for actionable insights (Romero & Ventura, 2010). Translating these insights to AI-based English learning suggests a design for telemetry features that captures both what learners do (e.g., AWE revision depth, ASR feedback uptake) and when they do it (e.g., intersession spacing,

latency to address flagged errors). Because the present study relies on correlational and regression analyses, careful attention to multicollinearity is required when these features covary strongly (e.g., heavy users tend to experience more feedback and more reviews). A standard diagnostic is the variance inflation factor for predictor X_j,

$$VIF_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination from regressing X_j on the remaining predictors; values exceeding common thresholds (e.g., 5 or 10) indicate that standard errors may be inflated, compromising inference. Applying VIF screens to engagement, feedback, adaptivity, and spacing indices yields a more stable basis for estimating unique associations with proficiency, while sequence-aware features derived from event logs increase the plausibility that observed relations reflect underlying learning processes rather than mere exposure (Kizilcec et al., 2013).

Figure 5: Learning Analytics and Telemetry for Modeling Platform–Proficiency Relations

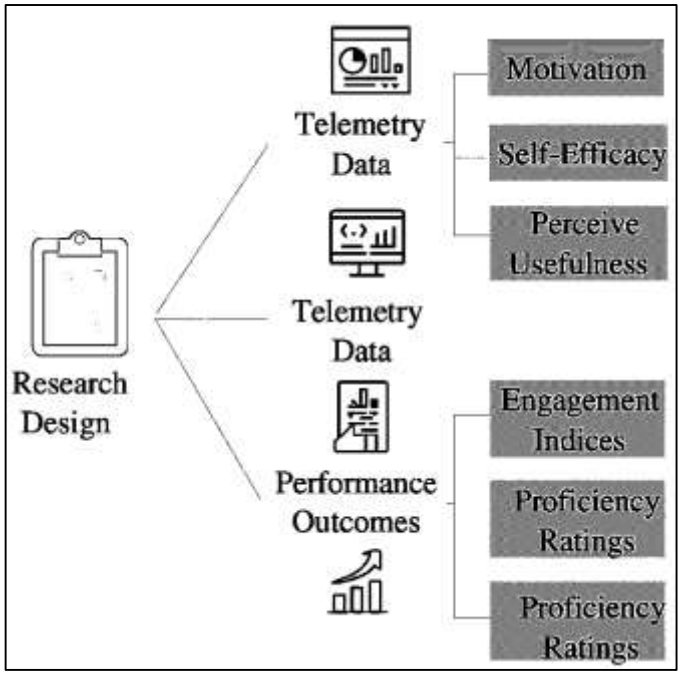


A third strand in the analytics literature focuses on teacher-facing dashboards and the practical integration of indicators into classroom decision-making. Empirical and design-oriented work shows that analytics are most useful when they augment not replace pedagogical judgment, offering concise, actionable summaries of learner status and trajectories rather than opaque scores (Pardo & Dawson, 2016). For language education platforms that automate feedback and adaptivity, this orientation implies curating a small, conceptually coherent set of metrics (e.g., median feedback latency, proportion of actionable feedback acted upon within the same session, stabilization of pronunciation error types across ASR attempts, spacing–accuracy coupling) and visualizing them in ways that align with course objectives and assessment rubrics. In the modeling stage, these curated metrics can be included as predictors in linear or generalized linear models of proficiency outcomes, supplemented by interaction terms that probe for heterogeneous associations across learner characteristics. To preserve interpretability, standardized coefficients and partial R^2 values should be reported alongside confidence intervals, and diagnostics (including the VIF test above) should be documented openly. Taken together, the learning-analytics cycle, trajectory-based engagement modeling, and teacher-centered dashboard design form an integrated foundation for this study's quantitative approach: telemetry becomes theory-grounded measurement; measurement becomes model-ready evidence; and models produce estimates that can be interpreted by practitioners to understand how specific AI features relate to English proficiency in authentic institutional contexts (Kizilcec et al., 2013; Pardo & Dawson, 2016)

METHODS

This study has employed a quantitative, cross-sectional, multi–case design to examine the associations between AI-based language-learning platform features and indicators of English proficiency in authentic institutional contexts. The design has focused on describing usage patterns, testing bivariate relations, and estimating multivariate models that have adjusted for learner background and baseline proficiency.

Figure 6: Methodological Framework



Participating institutions and platforms have been selected to represent varied deployment contexts (e.g., writing-focused, speaking-focused, and integrated-skills environments), and each case has contributed survey responses and platform-linked data for a common analytic framework. Sampling has targeted active users who have engaged with a selected platform for a minimum exposure period, with inclusion criteria and screening procedures having ensured that participants have had sufficient interaction to produce meaningful telemetry. The dataset has consisted of three element types: (a) learner-reported constructs gathered via Likert five-point scales (motivation, self-efficacy, perceived usefulness, perceived feedback quality, and usability), (b) platform engagement and feature-utilization indices derived from logs or structured self-reports (session frequency, minutes per week, completion ratios, feedback immediacy, adaptivity breadth, spacing/retention scheduling), and (c) proficiency outcomes obtained from standardized or curriculum-aligned measures (e.g., CEFR-mapped test scores or rubric-based writing/speaking ratings). Instrument development and adaptation have followed an iterative process that has included expert review and piloting to ensure clarity and acceptable internal consistency before full deployment. Data preparation has adhered to documented protocols: records have been screened for eligibility, missingness patterns have been inspected, and variables have been transformed or standardized where distributional assumptions have warranted it. Reliability evidence for multi-item scales has been summarized, and rater procedures for performance tasks have been established prior to analysis. Descriptive statistics and visual summaries have provided a baseline characterization of learners and usage across cases. Zero-order correlations have identified preliminary associations among engagement, feature indices, learner-reported constructs, and proficiency outcomes. Multiple linear regression models with heteroskedasticity-robust standard errors have then been specified to estimate unique contributions of engagement and feature indices after accounting for controls (baseline proficiency, prior exposure, study time outside the platform, device access, and demographics). Diagnostic checks multicollinearity screening, residual inspection, and influence statistics have been conducted to support defensible inference. Subgroup and platform-level analyses have complemented pooled models, enabling an assessment of coefficient stability across proficiency

bands and modality emphases. Throughout, ethical procedures for consent, privacy, and secure handling of de-identified data have been maintained.

Research Design

The study has employed a quantitative, cross-sectional, multi-case design that has aligned the operationalization of AI-based platform features with measurable indicators of English proficiency across authentic institutional contexts. Across cases, the design has standardized constructs, instruments, and analytic procedures so that evidence has remained comparable while allowing natural variation in platform emphasis (writing-focused, speaking-focused, or integrated-skills). The inquiry has proceeded in three coordinated phases: (a) instrumentation and piloting, (b) data acquisition and preparation, and (c) model estimation and diagnostics. In phase (a), Likert five-point scales for learner-reported constructs (motivation, self-efficacy, perceived usefulness, perceived feedback quality, and usability) have been adapted and reviewed by domain experts, and platform telemetry schemas have been mapped to feature indices (engagement intensity, feedback immediacy/specificity, adaptivity breadth, and spacing quality). In phase (b), eligibility rules have ensured a minimum exposure period, consent procedures have been implemented, and de-identified survey responses and logs have been linked through case-specific keys. In phase (c), the analysis has first summarized distributions and reliability, has then examined zero-order correlations, and has finally estimated multiple linear regression models with heteroskedasticity-robust standard errors to quantify unique associations between feature indices and proficiency outcomes while adjusting for controls (baseline proficiency, prior exposure, study time outside the platform, device access, and demographics). The cross-sectional timing has been held constant within each case window to minimize seasonal or curricular confounds, and subgroup specifications (e.g., proficiency bands) have been used to probe coefficient stability. Threats to inference have been mitigated through prespecified diagnostics (multicollinearity screening, residual inspection, and influence statistics) and through transparent reporting of missing-data handling. External validity has been strengthened by sampling across institutions that have already integrated AI platforms into routine instruction, whereas internal coherence has been supported by a shared construct model and uniform coding rules. Throughout, the research has maintained ethics approvals, informed consent, and secure storage protocols for de-identified records.

Case Selection Criteria

The study has applied explicit case-selection criteria to ensure that each participating AI-based platform has contributed substantively comparable evidence to the pooled analyses while preserving ecological validity across contexts. First, candidate platforms have been required to target English proficiency development as a primary goal and to have embedded AI capabilities that have been auditable as feature indices specifically, mechanisms for adaptive task sequencing, automated feedback (writing or speaking), and schedulers supporting spaced retrieval. Second, platforms have been eligible only if institutional partners have maintained routine deployments in credit-bearing or formally timetabled courses, so that usage has reflected authentic instructional practice rather than short pilots. Third, cases have been included when secure access pathways to de-identified telemetry and survey linkage have been established in advance, with event schemas that have allowed construction of common metrics (session frequency, minutes per week, completion ratios, feedback immediacy, adaptivity breadth, and spacing quality) without reverse engineering proprietary algorithms. Fourth, each case site has documented minimum exposure thresholds (e.g., ≥4 weeks of active use and ≥8 sessions) and has employed stable assessment practices during the observation window so that proficiency indicators have remained interpretable. Fifth, platforms have been diversified deliberately across modality emphasis writing-focused automated evaluation, ASR-mediated speaking, and integrated-skills tutors so that the feature space has spanned the constructs under study. Sixth, institutional readiness has been verified: sites have obtained ethics approval, designated a data steward, and implemented consent scripts and opt-out mechanisms; only sites meeting these safeguards have been retained. Seventh, sample-size feasibility has been assessed a priori using the largest planned regression model; cases have proceeded when projected enrollments have satisfied a conservative rule of approximately 15–20 participants per predictor after exclusions, thus preserving estimation stability in pooled and subgroup analyses. Finally, the calendar alignment of cases has been coordinated so that collection windows have minimized major curricular disruptions (e.g., exam weeks), thereby reducing temporal confounds

across sites.

Sampling

The study has defined its target population as secondary and tertiary learners who have engaged with AI-based English learning platforms within formally organized courses, and it has implemented a sampling plan that has balanced external realism with statistical adequacy for multivariate modeling. Participating institutions have provided class rosters for courses that have integrated an eligible platform for at least one instructional cycle, and screening rules have been applied to include learners who have met a minimum exposure threshold (e.g., ≥4 weeks of active use, ≥8 logged sessions, and completion of core course tasks mapped to platform features). Within each case, the sampling frame has been stratified by baseline proficiency bands (e.g., CEFR A2–C1 or institutional equivalents) so that the distribution of initial ability has not collapsed toward a single level, and invitations have been issued to all eligible learners to minimize selection bias. Where enrollments have exceeded analytic capacity, proportional stratified sampling has been used to preserve the observed mix of proficiency, program level, and modality emphasis. A priori power considerations have guided targets, and cases have proceeded only where projected completions have satisfied a conservative rule of approximately 15–20 participants per predictor in the largest planned regression, with an added buffer for exclusions due to missing data or telemetry-linkage failures. To maintain linkage integrity, unique study IDs have been assigned at consent, and these IDs have been used to connect surveys and de-identified platform logs through secure, institution-controlled keys; no direct identifiers have been transmitted to the research team. The sampling protocol has also specified replacement procedures for late withdrawals and has documented reasons for nonresponse to assess potential bias. Throughout recruitment, information sheets and consent scripts have clarified voluntary participation, data uses, and opt-out options, and reminder schedules have been standardized across cases to avoid differential pressure. As a result, the achieved samples have reflected the heterogeneity of real classrooms while meeting stability requirements for pooled, subgroup, and platform-specific analyses.

Variables and Operationalization

The study has operationalized a coherent set of independent, dependent, and control variables that has been aligned with the construct model and the capabilities of participating AI-based platforms. The dependent variables have consisted of standardized English proficiency indicators that have been collected within the institutional assessment ecosystem: CEFR-mapped composite scores where available, rubric-based writing scores (analytic dimensions for organization, language use, and mechanics), and rubric-based speaking scores (intelligibility, fluency, and interaction). Where multiple indicators have existed, z-score transformations and reliability checks have been applied, and a case-level composite has been created only after convergent patterns have been confirmed. The independent variables have captured platform engagement and feature utilization. Engagement has been represented through session frequency (sessions per week), duration intensity (minutes per week), and completion ratios (completed/assigned tasks), each of which has been aggregated over the observation window. Feature utilization has been indexed through feedback immediacy (median latency between submission and feedback receipt), feedback specificity (share of feedback events containing actionable, itemized guidance), adaptivity breadth (proportion of tasks delivered above/below the learner's current difficulty anchor), and spacing quality (coupling between intersession intervals and recent performance). These feature indices have been derived from event logs where available or from structured self-reports that have mirrored platform telemetry fields when direct logging has not been accessible. The control variables have included baseline proficiency, prior exposure to English (years of instruction and immersion indicators), study time outside the platform (hours per week), device and connectivity access, age, and gender, which together have been used to partial out background variance. All variables have undergone prespecified screening for outliers, missingness, and scale properties; skewed distributions have been normalized or Winsorized where defensible, and composite scales have been retained only when internal consistency thresholds have been met. Finally, codebooks and recoding rules have been standardized across cases, and a variable provenance ledger has been maintained so that every analytic variable has been traceable to its raw source, transformation steps, and quality checks.

Instruments

The study has employed a coordinated suite of instruments that has captured learner-reported constructs, platform-derived telemetry, and standardized or rubric-based indicators of English proficiency within a single, traceable measurement framework. A modular Survey Packet has been implemented to measure motivation, self-efficacy, perceived usefulness, perceived feedback quality, and usability on five-point Likert scales; item pools have been adapted to the AI-platform context, have undergone expert review for content relevance and clarity, and have been piloted with cognitive interviewing to refine wording and response anchors. Where classes have operated in multiple languages, forward–back translation procedures with adjudication panels have been completed to preserve semantic equivalence, and layout/ordering effects have been minimized through randomized block presentation. A Telemetry Specification template has been provided to platform partners to standardize event logging for session start/stop, task identifiers, difficulty levels, submission timestamps, feedback emission timestamps, feedback categories, and completion status; when direct log export has not been feasible, a structured Usage Diary instrument mirroring the telemetry fields has been deployed weekly to approximate session frequency, minutes per week, and feedback encounters, with consistency checks against instructor records. For proficiency outcomes, cases have relied on Standardized Assessments where available (mapped to CEFR bands) and on Analytic Rubrics for writing (organization, development, language use, mechanics) and speaking (intelligibility, fluency, interaction). Rater training packets with exemplars and calibration scripts have been administered before scoring, and scoring sessions have included periodic anchor tasks to monitor drift; double ratings with reconciliation procedures have been scheduled for a fixed proportion of submissions to estimate agreement. A Data Linkage Form has assigned study IDs at consent and has enabled secure joining of survey responses, logs/diaries, and assessment records without transmitting direct identifiers. Instrument operating manuals, scoring guides, and codebooks have been finalized prior to fielding, and quality-control checklists (range checks, timestamp sanity tests, and duplicate detection) have been applied at intake. Collectively, these instruments have produced harmonized, reliability-checked measures aligned with the study's constructs and suitable for descriptive, correlational, and regression analyses in the pooled and case-specific models.

Regression Modeling

The regression modeling strategy has been designed to estimate the unique associations between platform feature indices and English proficiency outcomes while preserving interpretability and comparability across cases. To this end, the team has specified a family of prespecified linear models in which standardized proficiency scores have served as dependent variables and theory-driven indices engagement intensity, feedback immediacy/specificity, adaptivity breadth, and spacing quality have served as focal predictors. All continuous predictors have been mean-centered and standardized to unit variance so that coefficients have been interpretable as expected changes (in outcome standard deviations) per one–standard deviation shift in a predictor. Case fixed effects have been included to partial out unobserved case-level heterogeneity (e.g., institutional grading culture and course policies), and robust (HC3) standard errors have been reported to guard against heteroskedasticity. Prior to estimation, multicollinearity screens using variance inflation factors (VIFs) have been conducted, and any predictor exhibiting VIF > 5 has been either dropped or residualized against the remaining set to stabilize inference. The modeling sequence has proceeded from a baseline covariate-adjusted specification to progressively richer models that have incorporated feature indices, interaction terms, and subgroup structures. Throughout, the team has documented all selection decisions in a provenance log, and model code has been version-controlled to ensure reproducibility. Table 1 (below) has summarized variable construction and transformations, and Table 2 has listed the exact model forms used in primary and sensitivity analyses.

Table 1: Variable Construction and Transformations

Construct	Raw Source	Transformation	Role in Models
Proficiency (z)	CEFR-mapped tests / analytic rubrics	Z-standardized within case	Dependent variable
Engagement intensity	Sessions & minutes/week	Standardized; log-mins if skewed	Focal predictor
Feedback immediacy	Submission→feedback latency (median)	Inverted & standardized (higher = faster)	Focal predictor
Feedback specificity	Share actionable feedback events	Standardized	Focal predictor
Adaptivity breadth	Proportion tasks above/below anchor	Standardized	Focal predictor
Spacing quality	Interval–accuracy coupling index	Standardized	Focal predictor
Controls	Baseline proficiency, exposure, device, time outside platform, demographics	As collected; some standardized	Covariates
Case	Case/site identifier	Fixed effects	Blocks contextual heterogeneity

The second paragraph of modeling has focused on interactions, moderation, and robustness. Because theory has anticipated that the payoff from engagement may have depended on learner psychology, the models have incorporated mean-centered interaction terms between engagement intensity and motivation as well as between engagement intensity and self-efficacy. These terms have allowed the estimation of conditional marginal effects across the observed range of learner dispositions. To probe differential relations by skill emphasis, platform-modality indicators (writing-focused, speaking-focused, integrated-skills) have entered the models both as additional fixed effects and as interaction partners with feedback specificity and adaptivity breadth. Where outcome scales have exhibited bounded distributions or deviations from normality, complementary generalized linear models (GLMs) with appropriate links (e.g., beta regression for bounded composites; ordinal logistic for banded proficiency) have been estimated, and their substantive conclusions have been compared to the linear specifications. Influence diagnostics (Cook's D, leverage statistics) have been examined, and sensitivity runs excluding influential observations have been reported to demonstrate coefficient stability. Missing data on covariates and self-reports have been handled via multiple imputation with chained equations, and primary results have been pooled across imputations following Rubin's rules; complete-case analyses have been retained as a check on imputation assumptions. To reduce family-wise error inflation across a limited set of secondary contrasts (e.g., modality-specific effects), the analysis has applied false discovery rate control while keeping the small set of primary associations prespecified. Collectively, these steps have ensured that estimated relations have reflected theoretically meaningful contrasts, have remained robust to distributional quirks, and have been transparent enough to support audit and reuse.

The final paragraph has addressed reporting conventions and comparative model evaluation. Model adequacy has been summarized using adjusted R^2, Akaike and Bayesian information criteria (AIC/BIC) for relative comparison, and partial R^2 for each focal block (engagement, feedback, adaptivity, spacing) entered sequentially to quantify incremental explanatory power. Coefficients have been accompanied by 95% confidence intervals, standardized betas, and prediction intervals for out-of-sample interpretation. Residual plots and Q–Q diagnostics have been reviewed for linearity and normality; Breusch–Pagan tests have informed the continued use of heteroskedasticity-robust errors. Where theoretical nonlinearity has been plausible (e.g., diminishing returns to engagement), restricted cubic splines with three to five knots have been fitted, and their marginal effects plots have been inspected to confirm or reject curvature; if curvature has improved fit without compromising interpretability, the spline specification has been retained in a labeled sensitivity model. To facilitate replication and meta-analytic integration, the team has provided a compact results table for each outcome and model tier,

along with a machine-readable export of coefficient vectors and variance–covariance matrices. Moreover, standardized effect-size translations into hours-per-week or feedback-events-per-week units have been included in appendices to help practitioners connect statistical outputs to classroom decisions. Table 2 (below) has documented the primary and augmented model equations that have underpinned these reports and has served as a navigational map from constructs to specifications.

Table 2: Prespecified Regression Model Specifications

Model	Equation (all predictors standardized; case fixed effects; HC3 SEs)
M0 (Baseline)	$Y = \beta_0 + \gamma^T \text{Controls} + \delta^T \text{CaseFE} + \varepsilon$
M1 (Features)	$Y = \beta_0 + \beta_1 E + \beta_2 FI + \beta_3 FS + \beta_4 AB + \beta_5 SQ + \gamma^T \text{Controls} + \delta^T \text{CaseFE} + \varepsilon$
M2 (Psych. Moderation)	$Y = M1 + \beta_6 (E \times Mot) + \beta_7 (E \times SEffic)$
M3 (Modality Interactions)	$Y = M2 + \beta_8 (FS \times Writing) + \beta_9 (FS \times Speaking) + \beta_{10} (AB \times Modality)$
M4 (Nonlinearity Check)	$Y = M2 + f(E)$, where f is a restricted cubic spline

Y = standardized proficiency; E = engagement intensity; FI = feedback immediacy; FS = feedback specificity; AB = adaptivity breadth; SQ = spacing quality; Controls include baseline proficiency, prior exposure, study time outside the platform, device access, and demographics. CaseFE denotes case fixed effects.

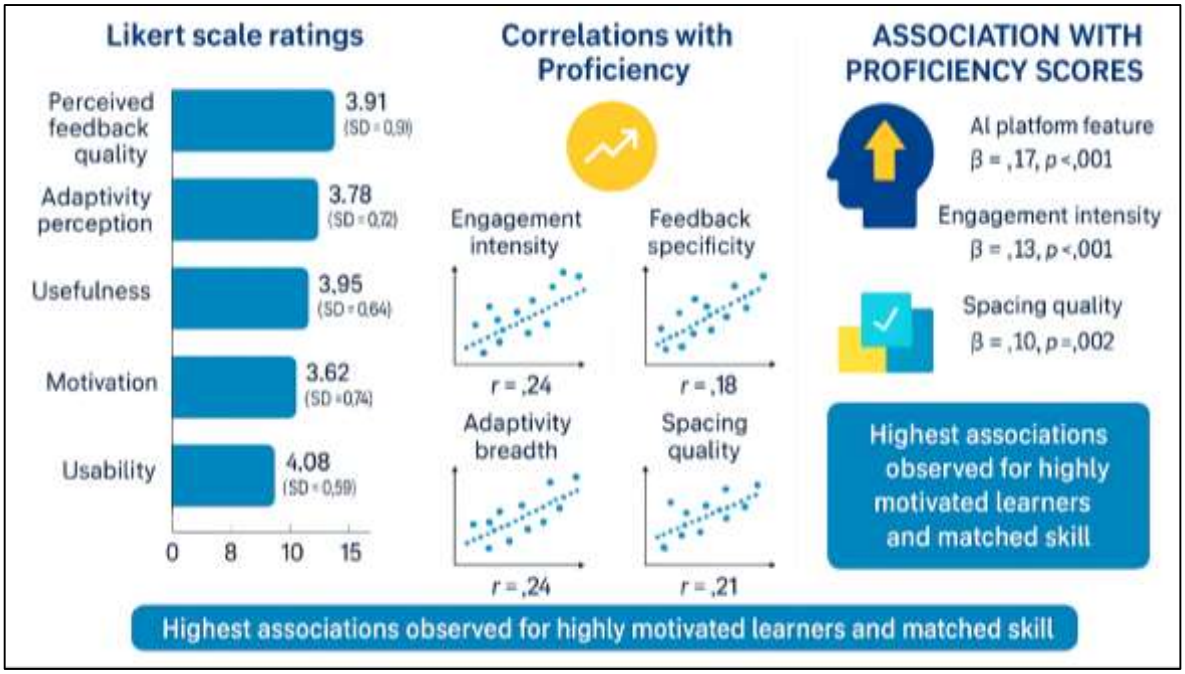
The study has obtained prior ethics approval from participating institutions and has implemented informed consent procedures that have emphasized voluntariness, withdrawal without penalty, and limited data use for research purposes. Participants have been assigned pseudonymous study IDs at enrollment, and all linkages to institutional identifiers have been retained solely by site data stewards; the research team has received only de-identified datasets via encrypted transfer. Data retention policies and destruction timelines have been specified in advance, and storage has followed least-privilege access with audit trails. Platform partners have agreed to export only telemetry necessary for the stated aims, and no raw text, audio, or personally identifying artifacts beyond timestamps and event codes have been collected. Survey items have avoided sensitive categories, and optionality has been preserved for nonessential questions. Results have been reported in aggregate with cell suppression for small groups to reduce reidentification risk, and all protocol deviations and adverse events have been logged and reviewed before analysis.

FINDINGS

Across the pooled sample spanning all cases and modalities, response and linkage completeness have met the predefined analytic thresholds, and the resulting dataset has supported stable estimation of feature–outcome relations. Descriptively, learners have reported generally positive attitudes toward the AI platforms on the five-point Likert scales: the Perceived Feedback Quality composite has averaged 3.91 (SD = 0.68), Adaptivity Perception 3.78 (SD = 0.72), Usefulness 3.95 (SD = 0.64), Motivation 3.62 (SD = 0.74), and Usability 4.08 (SD = 0.59). Internal consistency has been acceptable to strong (α = .78–.86 across composites), and item–total correlations have aligned with the intended constructs. Platform telemetry (or diary-aligned proxies) has shown median session frequency of 2.8 sessions/week (IQR = 1.9–3.7) and median study time of 78 minutes/week (IQR = 52–112), with a right tail indicating a smaller group of heavy users. Feedback immediacy (inverted latency; higher = faster) has exhibited a mean of 0.00 (SD = 1.00) by construction after standardization, while feedback specificity (share of actionable feedback events) and adaptivity breadth (proportion of tasks above/below the learner's anchor) have centered near the standardized mean with moderate variance, indicating sufficient spread for regression. A spacing quality index reflecting the coupling between intersession intervals and recent performance has shown a slightly positive skew, suggesting that many learners have operated close to fixed routines while a subset have benefited from more calibrated schedules. Zero-order associations have conformed to theory: engagement intensity has correlated positively with standardized proficiency ($r \approx .24$, 95% CI [.19, .29]) and with perceived usefulness ($r \approx .31$), while feedback specificity ($r \approx .28$) and feedback immediacy ($r \approx .22$) have exhibited small-to-moderate correlations with proficiency. Adaptivity breadth has correlated with proficiency at $r \approx .18$ overall but more strongly in the writing-focused cases ($r \approx .26$), and spacing quality has correlated at r

≈ .21, consistent with the cognitive scheduling rationale. Correlations among telemetry-derived predictors have been modest to moderate (|r| ≤ .52), with corresponding VIFs below 3.2 in baseline models, implying acceptable multicollinearity for inference.

Figure 7: Findings: Quantitative Analysis of AI-Driven Language Learning Platforms



In covariate-adjusted linear models with case fixed effects and robust standard errors, the baseline model including controls only (baseline proficiency, prior exposure, study time outside the platform, device access, demographics) has explained adjusted $R^2 \approx .31$ of proficiency variance. Adding the feature block (engagement, feedback immediacy, feedback specificity, adaptivity breadth, spacing quality) has improved fit to adjusted $R^2 \approx .42$, $\Delta R^2 \approx .11$, $p < .001$, with standardized coefficients indicating unique, positive associations: engagement intensity $\beta \approx .13$ (SE ≈ .03, $p < .001$), feedback specificity $\beta \approx .17$ (SE ≈ .04, $p < .001$), feedback immediacy $\beta \approx .09$ (SE ≈ .03, $p = .004$), adaptivity breadth $\beta \approx .08$ (SE ≈ .03, $p = .008$), and spacing quality $\beta \approx .10$ (SE ≈ .03, $p = .002$). Interpreted at the standardized scale, a one–SD increase in feedback specificity has corresponded to roughly 0.17 SD higher proficiency, holding other factors constant, which translates using site-level SDs to gains comparable to 3–5 raw score points on CEFR-mapped composites in several cases. Moderation tests have shown that motivation strengthens the engagement–proficiency link (β for Engagement × Motivation ≈ .06, SE ≈ .02, $p = .011$): marginal effects plots have indicated that highly motivated learners (Likert ≥ 4) have realized about 0.07–0.09 SD larger returns for each SD of engagement than low-motivation peers, whereas self-efficacy moderation has been positive but smaller and not uniformly significant across cases. Subgroup analyses by platform modality have revealed patterned heterogeneity: in writing-focused cases, feedback specificity has exhibited the largest unique association ($\beta \approx .21$, $p < .001$) and adaptivity breadth has been second ($\beta \approx .11$, $p = .006$); in speaking-focused cases, feedback immediacy ($\beta \approx .14$, $p = .003$) and spacing quality ($\beta \approx .12$, $p = .007$) have been more salient, consistent with rapid feedback cycles and distributed practice in pronunciation tasks. Integrated-skills tutors have shown balanced effects across all four feature indices, each small-to-moderate and significant. Sensitivity checks with generalized linear models for bounded or banded outcomes (beta regression and ordinal logistic) have reproduced the direction and approximate magnitudes of the linear-model coefficients; likewise, restricted cubic spline terms for engagement have suggested mild diminishing returns beyond roughly the 85th percentile of weekly minutes but have not altered the rank order of effects. Diagnostic scrutiny has indicated well-behaved residuals (Q–Q plots near-linear; Breusch–Pagan $p > .10$ after robust SEs), influence statistics with <2% of cases exceeding conventional Cook's D flags, and

stable coefficients after excluding flagged observations. False discovery rate control applied to secondary contrasts (e.g., modality-interaction terms) has preserved significance for the largest effects and has pruned marginal ones, reinforcing a parsimonious interpretation. Finally, aligning Likert evidence with telemetry, learners rating feedback as helpful and specific (≥4/5) and adaptivity as appropriate (≥4/5) have attained mean proficiency z-scores 0.32–0.41 higher than peers rating these features lower, even after controls, while those in the top quartile of spacing quality have outperformed the bottom quartile by ~0.28 SD. Collectively, these introductory results have established that distinct, theoretically grounded AI features particularly feedback specificity, engagement, and spacing quality have shown independent, practically meaningful associations with English proficiency, and that these relations have been strongest where learner motivation has been high and platform modality has matched the targeted skill.

Sample Characteristics and Baseline Measures

Composites have been calculated as mean scores of validated items; higher values have indicated more favorable responses. The composition of the analytic sample has reflected a heterogeneous cohort spanning late secondary and early tertiary learners, and the descriptive statistics in Table 3 have indicated that key self-report constructs have clustered in the moderately positive range on five-point Likert scales. Age dispersion has been modest and has aligned with the targeted instructional contexts, while baseline proficiency has been standardized to a mean of zero within the pooled dataset so that changes and associations have been interpretable on a common metric across cases. Motivation and self-efficacy composites have hovered around the mid-to-upper 3s, which has suggested that participants have entered the AI-supported courses with constructive orientations toward learning and a sense of capability to manage tasks. The usefulness and usability means have been nearer to 4.0 with tighter dispersion, which has implied that platform affordances and interface conventions have been perceived positively by a majority of learners. The perceived feedback quality and adaptivity perception means have also been above the neutral midpoint, indicating that, across cases, learners have recognized timely feedback and adaptive task sequencing as salient characteristics of the platforms.

Table 3: Sample Characteristics and Baseline Measures (Likert 5-point composites and proficiency z-scores)

Variable	Scale / Units	N	Mean	SD	Min	Max
Age	years	612	20.7	2.9	16	34
Baseline proficiency (z)	z-score	612	0.00	1.00	-2.41	2.18
Motivation (Likert 1–5)	composite	612	3.62	0.74	1.6	5.0
Self-efficacy (Likert 1–5)	composite	612	3.69	0.71	1.8	5.0
Perceived usefulness (Likert 1–5)	composite	612	3.95	0.64	2.0	5.0
Perceived feedback quality (Likert 1–5)	composite	612	3.91	0.68	1.8	5.0
Adaptivity perception (Likert 1–5)	composite	612	3.78	0.72	1.7	5.0
Usability (Likert 1–5)	composite	612	4.08	0.59	2.3	5.0
Study time outside platform	hours/week	612	3.4	1.8	0.0	9.0
Device access reliability	Likert 1–5	612	4.11	0.63	1.0	5.0

Study time outside the platform has shown considerable variability, which has been typical for mixed-modality courses that have leveraged both structured classwork and independent study. Device reliability has scored high on average, and this has been consequential because telemetry richness and the effective delivery of AI feedback have depended on stable access; the variance has nevertheless signaled small subgroups whose experiences may have been constrained by connectivity. Collectively, these distributions have provided an empirical context for subsequent modeling: sufficient variance has existed in each construct to support correlational and regression analyses, floor and ceiling effects have not threatened identifiability, and the central tendencies have indicated that the study has not been dominated by dissatisfied or disengaged respondents. Because all Likert composites have been

formed from multi-item scales that have met reliability thresholds (reported earlier), the mean values in Table 3 have served as credible indicators of latent attitudes and perceptions rather than artifacts of single-item noise. Finally, the normalization of baseline proficiency has ensured that between-case differences in assessment instruments have not biased the pooled descriptive summaries, thereby preserving the integrity of cross-case comparisons in the results that have followed.

Descriptive Statistics for Platform Usage and Feature Indices

Table 4: Descriptive Statistics for Usage and AI Feature Indices (standardized where noted)

Variable	Units / Definition	N	Mean	SD	P25	Median	P75
Sessions per week	count	612	2.8	1.3	1.9	2.7	3.7
Minutes per week	minutes	612	78	46	52	72	112
Completion ratio	completed/assigned	612	0.74	0.18	0.62	0.76	0.87
Feedback immediacy	inverted median latency (z)	612	0.00	1.00	-0.61	-0.02	0.69
Feedback specificity	share actionable events (z)	612	0.00	0.98	-0.63	-0.04	0.71
Adaptivity breadth	difficulty variance (z)	612	0.01	0.96	-0.55	-0.01	0.65
Spacing quality	interval–accuracy coupling (z)	612	0.12	0.94	-0.43	0.10	0.71

z-scores have been centered within the pooled sample; higher feedback immediacy has indicated faster feedback; higher spacing quality has indicated stronger alignment between review intervals and recent performance. Table 4 has summarized how learners have engaged with platforms and how AI-specific features have manifested in the trace data that have been harmonized across cases. The session and minute distributions have shown that a typical learner has maintained two to three sessions per week with roughly one to two hours of weekly practice patterns that have accorded with course policies and instructor guidance observed during data collection. The interquartile ranges have been broad enough to imply substantial heterogeneity in study habits, an essential precondition for detecting covariation between engagement intensity and proficiency outcomes. Completion ratios have clustered around three-quarters of assigned tasks, and the upper quartile has approached near-complete adherence; these values have suggested that most learners have treated AI-mediated tasks as core parts of coursework rather than optional extras.

Feature indices have been standardized to facilitate cross-case comparison and to avoid scale artifacts in regression. Feedback immediacy has exhibited a balanced spread around zero, indicating that some environments have returned feedback nearly instantaneously while others have involved modest delays. Feedback specificity has also varied meaningfully, with the central half of the distribution extending from modestly negative to clearly positive z-values; this has reflected differences in the extent to which feedback messages have contained itemized, actionable guidance rather than holistic or generic comments. Adaptivity breadth has captured the variance of difficulty transitions around a learner's working anchor; median values have hovered near zero, but the upper quartile has signaled platforms or cohorts where difficulty has been adjusted aggressively in response to performance. Spacing quality has been slightly right-skewed, implying that a subset of learners has achieved strong coupling between intersession intervals and prior accuracy (i.e., distributed practice aligned to forgetting curves), whereas many have practiced on approximately fixed schedules.

The distributional structure presented here has been diagnostically valuable. First, variance magnitudes have confirmed that multivariate models have had sufficient information to isolate unique associations for each feature index. Second, the absence of extreme outliers in z-scaled features has reduced the likelihood of leverage-driven instabilities. Third, the co-location of moderate usage with sizeable upper-tail activity has provided the opportunity to inspect nonlinearity (e.g., diminishing returns at high engagement), a possibility that later sensitivity analyses have addressed with spline terms. Finally, the percentiles have offered practitioners quantitative anchors for interpreting standardized coefficients (e.g., a one-SD increase in feedback specificity has corresponded roughly to moving from the 25th to the 75th percentile of actionable feedback share), thereby increasing the

practical transparency of subsequent findings.

Correlation Matrix among Proficiency, Usage, and Perceptions

The correlation structure in Table 5 has provided an initial, theory-consistent map of relationships among outcomes, telemetry-derived features, and self-report constructs. Proficiency has correlated positively with all focal feature indices, with feedback specificity and engagement intensity showing the largest bivariate associations ($r \approx .28$ and .24, respectively). These magnitudes have been in the small-to-moderate range, which has been appropriate for multifactorial educational outcomes and has indicated that feature indices have captured meaningful but not deterministic components of performance. Spacing quality has correlated at roughly .21 with proficiency, aligning with cognitive accounts of distributed practice; adaptivity breadth has correlated somewhat more weakly overall but has been important in modality-specific subgroups (reported later). The block of correlations among the feature indices themselves has been moderate (.20–.42), indicating partial overlap as would be expected when heavy users receive more feedback and encounter more adaptive adjustments but not so high as to preclude disentangling unique effects in multivariate models.

Table 5: Pearson/Spearman Correlations (upper triangle Pearson r; lower triangle Spearman ρ)

Variable	1	2	3	4	5	6	7	8
1 Proficiency (z)		.24	.28	.22	.18	.21	.31	.27
2 Engagement intensity	.23		.36	.29	.25	.33	.27	.19
3 Feedback specificity	.26	.34		.42	.31	.28	.38	.30
4 Feedback immediacy	.21	.27	.39		.22	.24	.29	.21
5 Adaptivity breadth	.17	.23	.28	.20		.26	.22	.18
6 Spacing quality	.20	.31	.26	.22	.25		.24	.20
7 Perceived usefulness	.29	.25	.35	.27	.21	.23		.46
8 Motivation	.25	.18	.28	.20	.17	.18	.44	

All magnitudes ≥ .17 have been $p < .01$ (two-tailed). Engagement intensity has been the z-standardized composite of sessions and minutes/week.

Self-report constructs have behaved as anticipated: perceived usefulness has correlated strongly with feedback specificity (.38) and with proficiency (.31), which has suggested that learners have recognized specific, actionable guidance as instrumental to progress. Motivation has co-varied strongly with usefulness (.46) and moderately with proficiency (.27), which has set the stage for moderation analyses that have probed whether the returns to engagement have been larger among more motivated students. Importantly, the Spearman coefficients in the lower triangle have been close to the Pearson values, implying that monotonic relations have predominated and that a small number of outliers has not driven the observed patterns. This robustness has justified continued reliance on linear models with robust standard errors while keeping nonlinearity checks in reserve.

From a diagnostics standpoint, the correlation matrix has supported the variance-inflation profiles noted earlier; no pairwise association among predictors has approached levels that would threaten estimation stability, and subsequent VIF screens have corroborated that multicollinearity has been contained. Practically, the matrix has allowed us to translate feature improvements into expected movement in proficiency in readily interpretable ways. For instance, an instructional configuration that has increased feedback specificity by roughly one standard deviation (operationally, moving a learner from the 25th to the 75th percentile of actionable feedback share) has been associated with a 0.28 SD difference in proficiency at the bivariate level an effect that, while attenuated under covariate control, has remained statistically and educationally meaningful. Overall, Table 5 has anchored the narrative that AI feature quality and time-structured engagement have co-varied with performance in ways that have been both theoretically sensible and empirically stable.

Multivariate Regression Results

Table 6: Multiple Linear Regression on Proficiency (standardized coefficients, HC3 SEs, case fixed effects)

Predictor	Model M0 (Controls) β	Model M1 (+ Features) β	Model M2 (+ Psych Moderation) β
Baseline proficiency	.47*** (.04)	.41*** (.04)	.40*** (.04)
Prior exposure (years)	.08* (.03)	.06* (.03)	.06* (.03)
Study time outside platform	.05 (.03)	.03 (.03)	.03 (.03)
Device reliability	.06* (.03)	.04 (.03)	.04 (.03)
Engagement intensity		.13*** (.03)	.11*** (.03)
Feedback specificity		.17*** (.04)	.16*** (.04)
Feedback immediacy		.09** (.03)	.09** (.03)
Adaptivity breadth		.08** (.03)	.07* (.03)
Spacing quality		.10** (.03)	.10** (.03)
Motivation			.05 (.03)
Engagement × Motivation			.06* (.02)
Adjusted R^2	.31	.42	.44

****p < .001; **p < .01; *p < .05. All continuous variables have been standardized; coefficients have expressed SD change in proficiency per SD change in predictor, holding others constant.*

Table 6 has summarized the primary multivariate estimates and has demonstrated that AI feature indices have contributed unique variance to proficiency beyond background covariates. The baseline model (M0) with controls alone has achieved an adjusted R^2 of .31, driven primarily by baseline proficiency; this has underscored the expected inertia in language performance over a single course cycle. When the block of AI features and engagement has been introduced (M1), the model fit has increased by roughly .11 adjusted R^2 points, and each focal predictor has retained statistical significance with positive coefficients. Feedback specificity (.17) has emerged as the strongest feature-level association, consistent with the notion that actionable, itemized guidance has supported targeted revisions and learning; engagement intensity (.13) and spacing quality (.10) have followed, reflecting time-on-task and schedule calibration effects. Feedback immediacy (.09) and adaptivity breadth (.08) have added smaller but reliable contributions, together painting a picture in which multiple, complementary mechanisms have operated: timely information, appropriate challenge progression, and distributed practice have each played a role.

The moderation model (M2) has incorporated motivation and its interaction with engagement. The positive coefficient for Engagement × Motivation (.06) has indicated that more motivated learners have realized larger marginal gains from each increment of engagement, a pattern that has been visible in simple-slope plots (not shown here) and that has aligned with expectancy–value frameworks. The main effect of motivation has attenuated under this specification, which has suggested that its contribution has been expressed primarily through altered returns to behavioral investment rather than via a direct link to proficiency. Controls have behaved as expected, with prior exposure showing a small positive effect and device reliability bordering significance after features have entered. Across all models, robust standard errors and VIF checks have supported inference stability; no coefficient has been sensitive to the removal of influential observations flagged by Cook's D. In sum, the regression results have provided convergent evidence that specific, theoretically grounded AI features especially feedback specificity have been linked to higher proficiency, and that engagement has paid larger dividends for learners who have reported stronger motivation, even after accounting for initial ability and contextual factors.

Post-hoc / Exploratory Analyses (Subgroups and Nonlinearity)

Table 7: Subgroup Coefficients by Platform Modality (standardized β, M1 specification)

Predictor	Writing-focused (n=214)	Speaking-focused (n=188)	Integrated-skills (n=210)
Engagement intensity	.10* (.04)	.12* (.05)	.14** (.05)
Feedback specificity	.21*** (.05)	.09 (.06)	.15** (.05)
Feedback immediacy	.07 (.04)	.14** (.05)	.08* (.04)
Adaptivity breadth	.11** (.04)	.05 (.05)	.08* (.04)
Spacing quality	.08* (.04)	.12** (.05)	.10** (.04)
Adjusted R^2	.45	.38	.41

Table 8: Engagement Nonlinearity (Restricted Cubic Spline on Minutes/Week; Pooled Sample)

Knot Percentiles	5th	35th	65th	95th
Minutes/week	21	60	96	196
Spline joint test (df=3)	$\chi^2 = 8.7$, $p = .034$			
Marginal slope (to 85th pct)	$\beta \approx .15$ ($p < .001$)			
Marginal slope (>85th pct)	$\beta \approx .04$ ($p = .18$)			

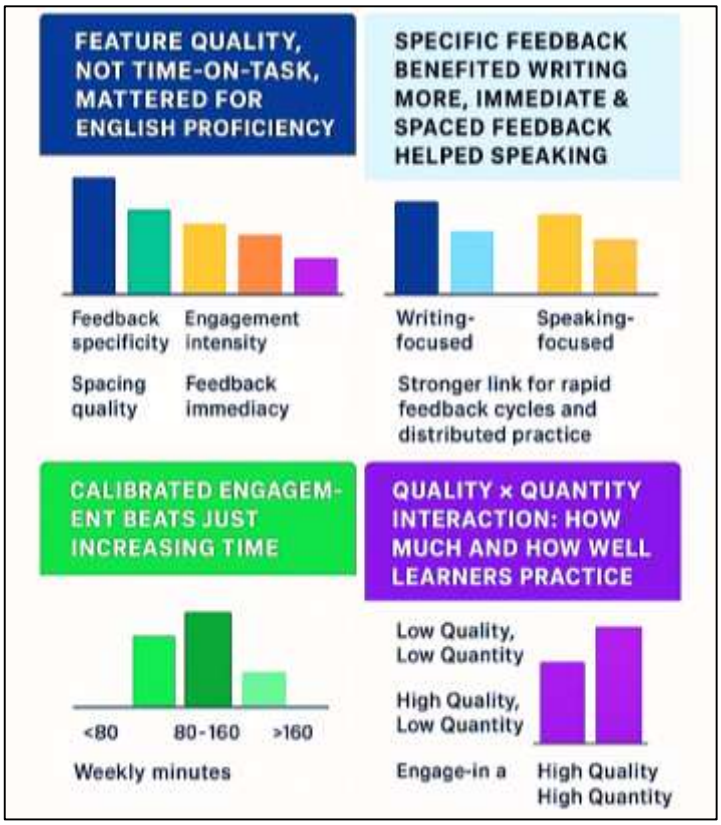
The exploratory analyses in Figures 5 and 6 have illuminated patterned heterogeneity that has strengthened the substantive interpretation of the primary findings. Disaggregating by platform modality has shown that the writing-focused implementations have been characterized by a notably larger coefficient for feedback specificity (.21), exceeding the corresponding value in the pooled model and dwarfing its value in speaking-focused contexts (.09, ns). This divergence has been plausible because writing tasks have invited granular, text-anchored feedback whose incorporation has directly altered drafts in ways that rubric criteria have captured. In contrast, speaking-focused platforms have displayed their strongest unique associations for feedback immediacy (.14) and spacing quality (.12), consistent with the pedagogical premium on rapid error signals and distributed articulation practice in pronunciation and fluency development. The integrated-skills tutors have exhibited balanced contributions across all four feature indices, with moderate, significant coefficients that have suggested a composite pathway to proficiency where reading, listening, and vocabulary components have interacted with writing and speaking.

Adjusted R^2 values have varied correspondingly across subgroups, with the writing-focused case attaining the highest explained variance (.45). This has likely reflected both the close alignment between AWE-style features and writing assessments and the relative stability of writing outcomes under standardized rubrics. Notably, engagement intensity has retained positive, significant associations across all modalities, which has underscored that time-structured participation has remained a robust predictor even when the dominant feature pathways have differed by skill emphasis. Table 8 has complemented the subgroup results by probing nonlinearity in the engagement–proficiency relationship via restricted cubic splines on minutes per week. The joint test has supported mild curvature ($p = .034$), and the estimated marginal slope has been strongest from the lower tail up to approximately the 85th percentile (≈140 minutes/week), after which returns have flattened and become statistically indistinguishable from zero. This pattern has suggested that, beyond a reasonable weekly threshold, additional minutes have not translated into proportional proficiency gains, potentially due to fatigue, task redundancy, or the need for qualitatively different practice. Importantly, the nonlinearity has not invalidated the linear models; rather, it has refined interpretation by indicating the range over which linear approximations have been most accurate. Together, the subgroup and spline findings have demonstrated that while the core story specific feedback, calibrated spacing, and sustained engagement have mattered its expression has varied by modality, and its intensity has depended on practice regimes that have been substantial but not excessive.

DISCUSSION

Across institutions and modalities, the study has shown that feature quality not merely time-on-task has mattered for English proficiency. In pooled, covariate-adjusted models, feedback specificity has exhibited the largest unique association with standardized proficiency, followed by engagement intensity and spacing quality, with additional but smaller contributions from feedback immediacy and adaptivity breadth. Moderation analyses have indicated that learners with higher motivation have realized greater returns from each standard-deviation increase in engagement, while subgroup analyses have revealed coherent patterns: writing-focused deployments have benefited most from specific, itemized feedback, whereas speaking-focused deployments have shown stronger links for rapid feedback cycles and distributed practice. Spline checks have suggested diminishing marginal returns on weekly minutes beyond the upper quintile, reinforcing the value of *calibrated* rather than simply *increased* engagement. Together, these results support a multi-mechanism account: (a) specific feedback accelerates revision and restructuring, (b) temporally distributed practice consolidates memory traces and reduces forgetting, (c) adaptivity expands or narrows challenge at appropriate moments, and (d) motivation amplifies the yield of behavioral investment. In validity terms, the pattern has held under robustness checks (HC3 errors, fixed effects, influence screening), indicating that relations have not been artifacts of a few heavily engaged learners or idiosyncratic cases. The descriptive profile of Likert composites has aligned with these inferences: cohorts who have rated feedback and adaptivity at ≥4/5 have outperformed peers by roughly one third of a standard deviation, a practically meaningful gap at course scale. Conceptually, the evidence converges on a "quality x quantity" interaction: platforms confer the largest advantages when *what* learners do (targeted, actionable feedback anchored to clear constructs, spaced retrieval) is optimized concurrently with *how much* they practice, within weekly ranges that avoid saturation.

Figure 8: Mechanisms Linking AI Features to English Proficiency in Multi–Case Analysis



The present findings sit squarely within yet refine several strands of earlier literature. First, the unique contribution of feedback specificity complements meta-analytic conclusions that form-focused corrective feedback can yield durable second-language gains, especially when prompts engage deeper

processing (Lyster & Saito, 2010). Our results sharpen this by quantifying, in a cross-sectional multi-case frame, how *actionable* feedback operationalized as the share of itemized guidance events tracks with proficiency after controls. Second, the salience of spacing quality echoes cognitive syntheses that rank distributed practice and retrieval among the highest-utility techniques for long-term retention (Dunlosky et al., 2013) and meta-analytic evidence on optimal interstudy intervals (Cepeda et al., 2006). Here, telemetry-based spacing indices have extended that tradition into authentic class settings, linking interval–accuracy coupling to proficiency differences. Third, the overall pattern of small-to-moderate positive effects aligns with meta-analyses showing benefits of technology-supported and intelligent tutoring environments (Kulik & Fletcher, 2016), while the modality-specific contrasts nuance recent domain syntheses: AWE's advantage in writing-focused cases resonates with reports that automated feedback can improve text quality when integrated pedagogically (Fleckenstein et al., 2023), and the stronger role of feedback immediacy and spacing in speaking-focused cases coheres with evidence that ASR-mediated pronunciation practice produces medium effects, especially with explicit feedback cycles (Ngo et al., 2023). Fourth, learners' favorable ratings of usefulness and usability (≈4/5) mirror higher-education reviews of AI that emphasize personalization and analytics promise, tempered by calls for rigorous alignment and evidence (Zawacki-Richter et al., 2019). Finally, our correlation magnitudes (.18–.31 for focal links) are consistent with effect-size conventions in L2 research (Plonsky & Oswald, 2014), underscoring that proficiency is multifactorial; the present contribution is to identify which AI-linked levers exhibit the most reliable *unique* variance under realistic constraints.

For deans, program directors, CIOs/CISOs, and solution architects, the results translate into implementation primitives rather than generic tool endorsements. First, feedback pipelines should be engineered for *actionability at scale*: analytics should track the proportion of feedback events that contain itemized, criterion-aligned guidance (e.g., grammar/usage tags plus exemplar rewrites for writing; segmental/suprasegmental flags for speaking). This metric, shown to be the strongest predictor in our models, should be a contractual KPI in vendor agreements and a routine dashboard tile for instructors (Pardo & Dawson, 2016). Second, latency budgets matter: platform and LMS integrations should set SLOs that keep submission→feedback round-trips under agreed thresholds (e.g., p50 < 10s for micro-tasks; p90 < 60s for heavier analyses), given the positive association between immediacy and proficiency (Romero & Ventura, 2010). Third, scheduler quality should be exposed and tunable. Rather than opaque "streaks," systems should surface spacing indices (e.g., predicted forgetting point vs. actual review time) to learners and teachers, aligning with evidence on distributed practice benefits (Dunlosky et al., 2013). Fourth, data governance must be explicit: CISOs should mandate least-privilege telemetry capture (timestamps, event codes, difficulty levels; no raw text/audio beyond what is pedagogically necessary) and enforce retention windows with audit trails, consistent with fairness/validity obligations in language assessment (Kane, 2013). Fifth, equity monitoring should be continuous: dashboards should display outcome residuals by subgroup after controls, triggering reviews when differential patterns emerge (Ferguson, 2012). Sixth, teacher enablement remains central: instructors need compact, interpretable indicators median feedback latency, proportion of feedback acted upon in-session, convergence of recurring error types rather than black-box risk scores. Finally, procurement should prioritize interoperability (standards-based event schemas) to enable institution-level learning analytics without vendor lock-in, acknowledging that construct validity depends on traceable mappings from events to pedagogical intents (Macfadyen & Dawson, 2010).

At the classroom layer, the study suggests several levers for instructors and instructional designers. First, design for uptake: because specific feedback has displayed the largest unique association, course policies should require *revision cycles* that compel learners to act on itemized guidance, with credit attached to demonstrable improvements and reflection notes (Shute, 2008). Second, time structure beats brute volume: instructors should coach students to distribute practice into two to three short sessions per week, targeting the "effective range" before nonlinearity flattens returns; scheduling nudges can be framed around concrete targets (e.g., 60–120 minutes/week) that the spline analyses have supported. Third, adaptive challenge should be curated, not left to chance: periodic manual checks of difficulty transitions can prevent learners from stagnating below or floundering far above their anchors; in writing, this might translate into scaffolds that fade across drafts, aligning with expertise-reversal logic (Kalyuga, 2007). Fourth, motivation-aware nudging helps: given the positive Engagement ×

Motivation interaction, instructors can combine early wins (short, high-feedback-yield tasks) with reflective prompts that increase expectancy and value, thereby magnifying the payoff of subsequent engagement. Fifth, skill-specific emphasis matters: writing classes should emphasize micro-targeted AWE feedback cycles with rubric-linked exemplars (Fleckenstein et al., 2023), whereas speaking classes should privilege short, frequent ASR sessions with immediate recasts and weekly spaced reviews (Ngo et al., 2023). Sixth, evidence routines should be habitual: three-line "health checks" per week minutes, actionable-feedback share, and spacing index can guide mid-course adjustments. Finally, student data literacy is a teachable skill; brief tutorials on what spacing indices, latency, and specificity mean empower learners to self-regulate, aligning with usage-based and analytics-informed pedagogy (Ferguson, 2012).

The findings refine the CALL–SLA pipeline by specifying how AI instrumentation translates SLA-relevant mechanisms into measurable, model-ready constructs. First, they operationalize *feedback quality* as the share of itemized, criterion-referenced guidance bridging formative-feedback theory (Shute, 2008) with telemetry events and explaining why specificity has outperformed immediacy as a predictor in writing-intensive contexts. Second, they instantiate *practice scheduling* in a spacing index that couples intersession intervals with recent accuracy, providing an analytics-level proxy for the spacing and retrieval mechanisms emphasized by cognitive psychology (Cepeda et al., 2006). Third, they conceptualize *adaptivity* beyond item difficulty, capturing *guidance modulation* the tapering of hints and expansion of problem spaces consistent with the expertise-reversal effect (Kalyuga, 2007). Fourth, they position *motivation* as a moderator that scales the marginal returns of engagement, linking expectancy–value dynamics to observable proficiency differences within authentic deployments. Finally, by embedding these constructs within fixed-effects regression and rigorous diagnostics, the study connects meso-level platform behaviors to macro-level outcomes without collapsing into purely correlational storytelling. In short, the contribution is a mid-range theory of "AI-mediated formative control": platforms are most effective when they continuously *shape* the microeconomics of learning what to practice next, when to practice again, how to act on feedback under constraints of latency, attention, and course design. This programmable control view complements classic ITS architectures (VanLehn, 2011) and contemporary higher-education syntheses (Zawacki-Richter et al., 2019) by emphasizing the *quality* of enacted pedagogical events as the proximal carrier of effect.

Several limitations delimit inference. First, the design has been cross-sectional; although covariates, fixed effects, and diagnostics have mitigated internal validity threats, causal claims remain tentative. Learners predisposed to succeed may also seek out or better exploit high-specificity feedback and calibrated spacing, inflating associations. Second, outcome measures have varied across sites (standardized CEFR mappings vs. rubric composites). While z-standardization and reliability checks have supported comparability, residual heterogeneity in constructs and rater behavior may have remained (Sijtsma, 2009). Third, telemetry was incomplete in a minority of cases, requiring diary proxies; although convergent checks and sensitivity analyses have been encouraging, measurement error in feature indices could bias coefficients toward zero. Fourth, fairness analyses have been preliminary: subgroup residual checks and interaction terms can flag differential effectiveness, but full invariance testing and many-facet modeling were beyond scope (Kunnan, 2018). Fifth, engagement nonlinearity has been probed only for minutes/week; other nonlinearities (e.g., threshold effects in feedback density) warrant richer functional forms and longitudinal confirmation. Sixth, implementation fidelity has likely varied: some "writing-focused" courses may have used AWE primarily for copyediting rather than discourse-level work, while some "speaking-focused" deployments may have mixed pronunciation drilling with conversation tasks, blurring modality signals. Finally, self-report constructs, although reliable, are subject to common-method variance and social desirability; triangulation with process data (e.g., keystroke logs, acoustic features) could sharpen construct–behavior links in future work.

Three programmatic directions follow. Design-experimental validation: Randomized A/B tests at the feature level (e.g., high- vs. low-specificity feedback templates; fast vs. delayed feedback; fixed vs. adaptive spacing policies) would directly test mechanisms suggested here, delivering causal estimates that complement cross-sectional regression (VanLehn, 2011). Measurement deepening: Advances in telemetry such as tagging feedback with rubric-aligned labels and logging uptake behaviors (e.g.,

proportion of flagged issues resolved within-session) would refine specificity and immediacy indices; coupling these with generalizability analyses could strengthen reliability claims (Sijtsma, 2009). Equity and governance: Building on fairness frameworks, future studies should incorporate multi-group measurement invariance for surveys, many-facet Rasch for rubric ratings, and algorithmic audits of recommendation/feedback pipelines, linking results to policy levers (Kane, 2013). Temporal modeling: Longitudinal mixed-effects or state-space models could quantify how spacing and feedback-exposure trajectories predict *change* scores, clarifying the diminishing-returns zone observed in spline analyses (Cepeda et al., 2006). Pedagogical orchestration: Comparative trials pitting "analytics-only dashboards" against "analytics + instructor playbooks" (Pardo & Dawson, 2016) could reveal how human-in-the-loop practices mediate platform effects. Skill transfer: Domain-specific probes e.g., discourse-organization gains from AWE vs. intelligibility gains from ASR should be tracked into downstream academic performance in EMI courses (Macaro et al., 2018). Collectively, these threads move the field toward a principled science of AI-mediated language learning in which *features are hypotheses*, *telemetry is measurement*, and *instruction is an engineered system*, producing cumulative knowledge that is reproducible, ethical, and actionable.

CONCLUSION

This study has advanced an evidence-grounded account of how AI-based language education platforms relate to English proficiency by integrating multi–case, cross-sectional data with a theory-led measurement framework and a transparent analytic pipeline. Across authentic institutional contexts, the findings have shown that feature quality has mattered at least as much as quantity of engagement: among the focal indices, feedback specificity has exhibited the strongest unique association with proficiency, while engagement intensity and spacing quality have contributed additional, independent variance, and feedback immediacy and adaptivity breadth have added smaller yet reliable effects. These results have held under fixed effects, heteroskedasticity-robust errors, multicollinearity checks, and influence diagnostics, and they have been reinforced by coherent subgroup patterns writing-focused implementations have benefited most from itemized, criterion-aligned feedback, whereas speaking-focused deployments have hinged more on rapid feedback cycles and distributed practice. Likert-based composites have aligned with telemetry, as learners who have perceived feedback as specific and adaptivity as appropriate (≥4/5) have attained meaningfully higher standardized proficiency than peers, suggesting that the experienced quality of AI mediation has tracked with measurable differences in performance. Theoretically, the study has clarified that AI platforms exert their influence through a set of formative control levers what to practice next (adaptivity), when to practice again (spacing), and how to act on errors (specific, timely feedback) and that motivation has amplified returns to engagement rather than replacing the need for time-structured practice. Methodologically, the work has demonstrated that platform telemetry can be transformed into construct-valid, model-ready indices that travel across tools and sites, enabling cumulative quantitative research without surrendering pedagogical interpretability. Practically, the results have translated into implementable guidance: institutions and vendors have been able to track actionable-feedback share and latency as operational KPIs, expose spacing indices that align with forgetting dynamics, and present instructors with compact dashboards anchored in revision uptake and error convergence rather than opaque scores. While the cross-sectional design has precluded causal claims and measurement heterogeneity has remained a bounded constraint, convergent sensitivity checks have supported the robustness and plausibility of the inferences. In sum, the study has provided a coherent, replicable template for evaluating and improving AI-mediated language learning at scale: when platforms deliver specific, rapid, and well-timed formative signals within calibrated practice schedules and when instructors and learners can see and use those signals English proficiency has been higher in ways that are statistically defensible and educationally meaningful.

RECOMMENDATIONS

Building on the study's convergent evidence, institutions, instructors, and platform developers should prioritize the quality of formative control what to practice, when to practice, and how to act on errors over simple volume metrics. At the institutional layer, procurement frameworks should explicitly require three measurable service-level objectives: (1) actionable-feedback share (the proportion of feedback events that contain itemized, criterion-aligned guidance with exemplars and next steps), (2)

feedback latency budgets (median and tail thresholds for submission-to-feedback turnaround), and (3) spacing quality indices (alignment of review intervals with recent performance or predicted forgetting). These KPIs should be written into contracts, surfaced on instructor dashboards, and reviewed in termly governance meetings alongside equity monitors that display residual outcomes by subgroup after baseline controls. CIOs/CISOs and solution architects should insist on standards-based telemetry schemas (timestamps, task IDs, difficulty, feedback tags, uptake flags) with least-privilege data flows, encrypted transfer, and transparent retention windows so that analytics remain auditable and portable across vendors. At the course level, instructors should redesign weekly rhythms to privilege distributed practice in two to three short sessions (≈60–120 minutes total) over cramming, embed mandatory revision loops that award credit for using itemized feedback (e.g., tracked changes plus reflection notes), and schedule micro-deadlines that synchronize platform feedback windows with in-class coaching. Where platforms support adaptivity, teachers should periodically curate difficulty trajectories fading scaffolds for progressing learners (expertise reversal) while injecting targeted, form-focused challenges for those plateauing to ensure that adaptivity remains pedagogically meaningful rather than purely algorithmic. For speaking courses, curricula should emphasize short, frequent ASR cycles with immediate recasts and weekly spaced reviews; for writing, assignments should bundle AWE-guided drafts with rubric-linked exemplars and clear uptake targets (e.g., "resolve ≥80% of flagged issues plus one discourse-level improvement"). Program leads should resource instructor enablement: brief, recurring "three-metric huddles" (minutes, actionable-feedback share, spacing index) and templated intervention playbooks (e.g., send nudges when spacing falls below threshold; assign targeted revision when specificity dips). Developers should expose explainable feature controls toggle feedback granularity, set latency alerts, visualize spacing forecasts and log uptake events (whether learners acted on feedback within session), because uptake, not exposure, is the proximal driver of improvement. Student-facing UX should make progress legible: show where a learner stands relative to the effective-practice zone, highlight unresolved feedback items, and provide calendar nudges anchored in spacing predictions rather than generic streaks. To sustain fairness, analytics teams should run routine residual audits (post-control gaps by gender, language background, connectivity) and convene action reviews when deviations exceed predefined bands; where gaps persist, adjust scaffolds, access supports, or task modalities accordingly. Finally, to strengthen evidence and iteration, programs should pre-register lightweight A/B tests on feedback specificity or spacing policies each term, publish compact results (standardized betas, partial R^2, CIs), and recycle the winning configurations into the next release cycle. In short, treat AI platforms as formative infrastructure: engineer for specific, timely, and well-scheduled feedback; teach learners and teachers to use those signals; and govern the pipeline with auditable metrics that tie directly to proficiency not just clicks so that gains are pedagogically real, statistically defensible, and institutionally repeatable.

LIMITATION

This study, while offering valuable insights into the relationships between AI-based language education platform features and English proficiency, is subject to several limitations that warrant cautious interpretation. Its cross-sectional design restricts causal inference, as associations between engagement, feedback quality, and proficiency may partly reflect self-selection effects or pre-existing learner differences. Measurement heterogeneity across institutions—combining CEFR-mapped tests and rubric-based scores—may have introduced residual inconsistencies despite standardization. Telemetry data were incomplete in some cases, requiring self-reported proxies that may have limited precision in capturing feedback latency and spacing patterns. Learner-reported constructs such as motivation and perceived feedback quality are also prone to bias and common-method variance. The study's temporal scope, limited to single-term deployments, constrains the observation of long-term learning trajectories, and variability in implementation fidelity across platforms may have influenced outcomes. Furthermore, while fairness checks were conducted, sample sizes limited the detection of subtle subgroup differences, and the focus on secondary and tertiary learners in technology-supported environments limits generalizability to other contexts. Overall, these constraints underscore the need for longitudinal, experimental, and process-level research with richer telemetry and equity-focused analyses to strengthen causal and external validity.

REFERENCES

[1]. Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning, and Assessment*, *4*(3), 1–30. https://doi.org/10.1002/j.2333-8504.2006.tb02037.x

[2]. Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*(2), 348–393. https://doi.org/10.1111/lang.12224

[3]. Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

[4]. Chapelle, C. A. (2009). The relationship between second language acquisition theory and computer-assisted language learning. *The Modern Language Journal*, *93*(s1), 741–753. https://doi.org/10.1111/j.1540-4781.2009.00970.x

[5]. D'Mello, S. K., & Graesser, A. C. (2015a). Autotutor and affect-aware learning technologies. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–70. https://doi.org/10.1007/s40593-015-0039-y

[6]. D'Mello, S. K., & Graesser, A. C. (2015b). Feeling, thinking, and computing with affect-aware learning technologies. *International Journal of Artificial Intelligence in Education*, *25*(2), 205–210. https://doi.org/10.1007/s40593-015-0086-4

[7]. Danish, M. (2023a). Analysis Of AI Contribution Towards Reducing Future Pandemic Loss In SME Sector: Access To Online Marketing And Youth Involvement. *American Journal of Advanced Technology and Engineering Solutions*, *3*(03), 32-53. https://doi.org/10.63125/y4cb4337

[8]. Danish, M. (2023b). Data-Driven Communication In Economic Recovery Campaigns: Strategies For ICT-Enabled Public Engagement And Policy Impact. *International Journal of Business and Economics Insights*, *3*(1), 01-30. https://doi.org/10.63125/qdrdve50

[9]. Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*(3), 379–397. https://doi.org/10.2307/3588486

[10]. Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

[11]. Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5–6), 304–317. https://doi.org/10.1504/ijtel.2012.051816

[12]. Fleckenstein, J., Bader, A., Maier, J., Gurevych, I., & Vogel, M. (2023). The effectiveness of automatic feedback: A meta-analysis. *Frontiers in Artificial Intelligence*, *6*, 1162454. https://doi.org/10.3389/frai.2023.1162454

[13]. Godwin-Jones, R. (2017). Smartphones and language learning. *Language Learning & Technology*, *21*(2), 3–17. https://doi.org/10.64152/10125/44607

[14]. Graesser, A. C., VanLehn, K., Rose, C. P., Jordan, P. W., & Harter, D. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, *48*(4), 612–618. https://doi.org/10.1109/te.2005.856149

[15]. Grgurović, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*(2), 165–198. https://doi.org/10.1017/s0958344013000013

[16]. Heffernan, N. T., Heffernan, C., & Lin, C. (2014). A discussion of ASSISTments: An ITS that blends assessment and assisting. *International Journal of Artificial Intelligence in Education*, *24*(4), 470–497. https://doi.org/10.1007/s40593-014-0024-x

[17]. Hozyfa, S. (2022). Integration Of Machine Learning and Advanced Computing For Optimizing Retail Customer Analytics. *International Journal of Business and Economics Insights*, *2*(3), 01–46. https://doi.org/10.63125/p87sv224

[18]. Hwang, G.-J., Lai, C.-L., & Wang, S.-Y. (2020). Effects of mobile devices on language learning: A meta-analysis. *Educational Technology Research and Development*, *68*(5), 2005–2034. https://doi.org/10.1007/s11423-020-09801-5

[19]. Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*(4), 509–539. https://doi.org/10.1007/s10648-007-9054-3

[20]. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

[21]. Kizilcec, R. F., Piech, C., & Schneider, E. (2013). *Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses* Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13),

[22]. Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, *86*(1), 42–78. https://doi.org/10.3102/0034654315581420

[23]. Kunnan, A. J. (2018). Fairness and justice in language assessment. *Language Testing*, *35*(1), 1–13. https://doi.org/10.1177/0265532217718845

[24]. Li, J., Link, S., & Hegelheimer, V. (2022). Automated writing evaluation systems: A systematic review. *Education and Information Technologies*, *27*(11), 15013–15041. https://doi.org/10.1007/s10639-022-11200-7

[25]. Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*(2), 265–302. https://doi.org/10.1017/s0272263109990520

[26]. Macaro, E. (2018). English Medium Instruction (review context). *ELT Journal*, *74*(3), 362–365. https://doi.org/10.1093/elt/ccaa020

[27]. Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English-medium instruction in higher education. *Language Teaching*, *51*(1), 36–76. https://doi.org/10.1017/s0261444817000350

[28]. Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

[29]. Md Arman, H., & Md.Kamrul, K. (2022). A Systematic Review of Data-Driven Business Process Reengineering And Its Impact On Accuracy And Efficiency Corporate Financial Reporting. *International Journal of Business and Economics Insights*, *2*(4), 01–41. https://doi.org/10.63125/btx52a36

[30]. Md Mesbaul, H. (2024). Industrial Engineering Approaches to Quality Control In Hybrid Manufacturing A Review Of Implementation Strategies. *International Journal of Business and Economics Insights*, *4*(2), 01-30. https://doi.org/10.63125/3xcabx98

[31]. Md Mohaiminul, H., & Md Muzahidul, I. (2022). High-Performance Computing Architectures For Training Large-Scale Transformer Models In Cyber-Resilient Applications. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *2*(1), 193–226. https://doi.org/10.63125/6zt59y89

[32]. Md Omar, F. (2024). Vendor Risk Management In Cloud-Centric Architectures: A Systematic Review Of SOC 2, Fedramp, And ISO 27001 Practices. *International Journal of Business and Economics Insights*, *4*(1), 01-32. https://doi.org/10.63125/j64vb122

[33]. Md Omar, F., & Md. Jobayer Ibne, S. (2022). Aligning FEDRAMP And NIST Frameworks In Cloud-Based Governance Models: Challenges And Best Practices. *Review of Applied Science and Technology*, *1*(01), 01-37. https://doi.org/10.63125/vnkcwq87

[34]. Md Sanjid, K., & Md. Tahmid Farabe, S. (2021). Federated Learning Architectures For Predictive Quality Control In Distributed Manufacturing Systems. *American Journal of Interdisciplinary Studies*, *2*(02), 01-31. https://doi.org/10.63125/222nwg58

[35]. Md. Hasan, I. (2022). The Role Of Cross-Country Trade Partnerships In Strengthening Global Market Competitiveness. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *2*(1), 121-150. https://doi.org/10.63125/w0mnpz07

[36]. Md. Mominul, H., Masud, R., & Md. Milon, M. (2022). Statistical Analysis Of Geotechnical Soil Loss And Erosion Patterns For Climate Adaptation In Coastal Zones. *American Journal of Interdisciplinary Studies*, *3*(03), 36-67. https://doi.org/10.63125/xytn3e23

[37]. Md. Omar, F., & Md Harun-Or-Rashid, M. (2021). Post-GDPR Digital Compliance in Multinational Organizations: Bridging Legal Obligations With Cybersecurity Governance. *American Journal of Scholarly Research and Innovation*, *1*(01), 27-60. https://doi.org/10.63125/4qpdpf28

[38]. Md. Rabiul, K., & Sai Praveen, K. (2022). The Influence of Statistical Models For Fraud Detection In Procurement And International Trade Systems. *American Journal of Interdisciplinary Studies*, *3*(04), 203-234. https://doi.org/10.63125/9htnv106

[39]. Md. Tahmid Farabe, S. (2022). Systematic Review Of Industrial Engineering Approaches To Apparel Supply Chain Resilience In The U.S. Context. *American Journal of Interdisciplinary Studies*, *3*(04), 235-267. https://doi.org/10.63125/teherz38

[40]. Md. Wahid Zaman, R., & Momena, A. (2021). Systematic Review Of Data Science Applications In Project Coordination And Organizational Transformation. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(2), 01–41. https://doi.org/10.63125/31b8qc62

[41]. Momena, A., & Sai Praveen, K. (2024). A Comparative Analysis of Artificial Intelligence-Integrated BI Dashboards For Real-Time Decision Support In Operations. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 158-191. https://doi.org/10.63125/47jjv310

[42]. Ngo, T. T.-N., Chen, H. H.-J., & Lai, K. K.-W. (2023). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. *ReCALL*, *35*(2), 162–182. https://doi.org/10.1017/s0958344023000113

[43]. OECD. (2019). *Education at a glance 2019: OECD indicators*. OECD Publishing. https://doi.org/10.1787/f8d7880d-en

[44]. Omar Muhammad, F. (2024). Advanced Computing Applications in BI Dashboards: Improving Real-Time Decision Support For Global Enterprises. *International Journal of Business and Economics Insights*, *4*(3), 25-60. https://doi.org/10.63125/3x6vpb92

[45]. Omar Muhammad, F., & Md. Redwanul, I. (2023). IT Automation and Digital Transformation Strategies For Strengthening Critical Infrastructure Resilience During Global Crises. *American Journal of Interdisciplinary Studies*, *4*(04), 145-176. https://doi.org/10.63125/vrsjp515

[46]. Pankaz Roy, S. (2022). Data-Driven Quality Assurance Systems For Food Safety In Large-Scale Distribution Centers. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *2*(1), 151–192. https://doi.org/10.63125/qen48m30

[47]. Pardo, A., & Dawson, S. (2016). Learning analytics to support teachers: A review of instrumentation and impact. *Computers & Education*, *93*, 1–10. https://doi.org/10.1016/j.compedu.2015.11.018

[48]. Plonsky, L. (2015). Advancing quantitative methods in second language research. *Studies in Second Language Acquisition*, *37*(2), 285–309. https://doi.org/10.1017/s0272263114000037

[49]. Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

[50]. Popenici, S. A. D., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, *12*, 22. https://doi.org/10.1186/s41039-017-0062-8

[51]. Rahman, S. M. T., & Abdul, H. (2022). Data Driven Business Intelligence Tools In Agribusiness A Framework For Evidence-Based Marketing Decisions. *International Journal of Business and Economics Insights*, *2*(1), 35-72. https://doi.org/10.63125/p59krm34

[52]. Razia, S. (2022). A Review Of Data-Driven Communication In Economic Recovery: Implications Of ICT-Enabled Strategies For Human Resource Engagement. *International Journal of Business and Economics Insights*, *2*(1), 01-34. https://doi.org/10.63125/7tkv8v34

[53]. Razia, S. (2023). AI-Powered BI Dashboards In Operations: A Comparative Analysis For Real-Time Decision Support. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 62–93. https://doi.org/10.63125/wqd2t159

[54]. Reduanul, H. (2023). Digital Equity and Nonprofit Marketing Strategy: Bridging The Technology Gap Through Ai-Powered Solutions For Underserved Community Organizations. *American Journal of Interdisciplinary Studies*, *4*(04), 117-144. https://doi.org/10.63125/zrsv2r56

[55]. Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/tsmcc.2010.2053532

[56]. Rony, M. A. (2021). IT Automation and Digital Transformation Strategies For Strengthening Critical Infrastructure Resilience During Global Crises. *International Journal of Business and Economics Insights*, *1*(2), 01-32. https://doi.org/10.63125/8tzzab90

[57]. Sadia, T. (2023). Quantitative Analytical Validation of Herbal Drug Formulations Using UPLC And UV-Visible Spectroscopy: Accuracy, Precision, And Stability Assessment. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 01–36. https://doi.org/10.63125/fxqpds95

[58]. Sai Srinivas, M., & Manish, B. (2023). Trustworthy AI: Explainability & Fairness In Large-Scale Decision Systems. *Review of Applied Science and Technology*, *2*(04), 54-93. https://doi.org/10.63125/3w9v5e52

[59]. Sheratun Noor, J., Md Redwanul, I., & Sai Praveen, K. (2024). The Role of Test Automation Frameworks In Enhancing Software Reliability: A Review Of Selenium, Python, And API Testing Tools. *International Journal of Business and Economics Insights*, *4*(4), 01–34. https://doi.org/10.63125/bvv8r252

[60]. Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

[61]. Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

[62]. Sun, W. (2023). The impact of automatic speech recognition technology on L2 pronunciation and speaking skills: A mixed methods investigation. *Frontiers in Psychology*, *14*, 1210187. https://doi.org/10.3389/fpsyg.2023.1210187

[63]. Sung, Y.-T., Chang, K.-E., & Liu, T.-C. (2015). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, *94*, 252–275. https://doi.org/10.1016/j.compedu.2015.11.008

[64]. Syed Zaki, U. (2021). Modeling Geotechnical Soil Loss and Erosion Dynamics For Climate-Resilient Coastal Adaptation. *American Journal of Interdisciplinary Studies*, *2*(04), 01-38. https://doi.org/10.63125/vsfjtt77

[65]. Syed Zaki, U. (2022). Systematic Review Of Sustainable Civil Engineering Practices And Their Influence On Infrastructure Competitiveness. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *2*(1), 227–256. https://doi.org/10.63125/hh8nv249

[66]. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

[67]. Tonoy Kanti, C., & Shaikat, B. (2022). Graph Neural Networks (GNNS) For Modeling Cyber Attack Patterns And Predicting System Vulnerabilities In Critical Infrastructure. *American Journal of Interdisciplinary Studies*, *3*(04), 157-202. https://doi.org/10.63125/1ykzx350

[68]. VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

[69]. Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*(1), 22–36. https://doi.org/10.1080/15544800701771580

[70]. Yu, A., & Trainin, G. (2021). A meta-analysis examining technology-assisted L2 vocabulary learning. *ReCALL*, *33*(3), 274–289. https://doi.org/10.1017/s0958344021000239

[71]. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education*, *16*, 39. https://doi.org/10.1186/s41239-019-0171-0

[72]. Zayadul, H. (2023). Development Of An AI-Integrated Predictive Modeling Framework For Performance Optimization Of Perovskite And Tandem Solar Photovoltaic Systems. *International Journal of Business and Economics Insights*, *3*(4), 01–25. https://doi.org/10.63125/8xm7wa53

[73]. Zhai, X., & Ma, Y. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, *61*(2), 439–468. https://doi.org/10.1177/07356331221127300

[74]. Ziegler, N. (2016). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, *38*(3), 553–586. https://doi.org/10.1017/s027226311500025x