International Journal of Information and Education Technology, Vol. 16, No. 7, 2026

Multimodal Immersive Language Learning: A Framework Linking AR/VR Features with Language Skill Development

Saima Mehboob * and Alberto Fornasari

Department of Educational Sciences, Psychology, Communication, University of Bari Aldo Moro, Bari, Italy
Email: s.mehboob@phd.uniba.it (S.M.); alberto.fornasari@uniba.it (A.F.)
*Corresponding author
Manuscript received December 10, 2025; revised January 8, 2026; accepted March 16, 2026; published July 17, 2026

***Abstract*—This paper presents a Virtual Reality (VR)-based framework for evaluating English language proficiency in an immersive learning environment. Based on a dataset of 176,911 participants, we analyzed reading, listening, speaking, and pronunciation activities, as well as vocabulary acquisition, writing exercises, and grammar correction. The translation tasks and language-learning results are complemented by behavioral indicators, including gesture frequency, reaction time, interaction depth, and session length. Multivariate regression results indicated a moderate proficiency level and weak associations between behavioral engagement and language outcomes. Neither the device type nor interaction metrics had a significant effect on overall proficiency level, and mediation testing further revealed no indirect effects via response latency. These results suggest that VR language skill learning does not depend on the quantity of behavioral activity and highlight the significance of task design and meaningful engagement. This framework is designed to help build future adaptive, scalable language assessment systems using immersive technologies to support fair, data-driven English language learning.**

***Keywords*—English language proficiency, virtual reality learning, immersive assessment, behavioral analytics, gesture interaction, response latency, adaptive learning systems**

I. Introduction

Multimodal immersive language learning blends Augmented Reality (AR) / Virtual Reality (VR) / Mixed Reality (MR) technologies with educational theory to create highly interactive environments that engage learners not only visually and aurally but also through physical and sensorimotor participation; it requires learners to think physically and sensorially. This approach is transforming language instruction by providing authentic, meaningful content in immersive video-based contexts.

In recent years, immersive technologies such as AR and VR have emerged as powerful tools, providing students with authentic, real-world learning contexts. In language acquisition, these are new opportunities for situated practice, motivation, and multimodal interaction. The use of AR/VR in language learning builds on the growing demand for more engaging, efficient, and context-rich learning environments. Such tools place the learner in impactful, timely ways and provide experiences that are difficult to obtain through conventional methods of learning, such as experience-based or situated learning. VR, for instance, allows learners to reinforce language concepts through gestures, movement, and spatial navigation [1, 2].

Previous studies report that AR/VR environments enhance vocabulary, speaking, and writing skills, while also reducing anxiety and increasing motivation [3]. While language learning technology enhances user satisfaction and engagement, challenges remain regarding technical expertise, collaboration, instructional design, and implementation [4]. Multimodal Immersive Virtual Reality (MIVR) offers natural, authentic learning experiences but also presents challenges related to accessibility, cognitive load, and personalization [5].

While prior research confirms the general efficacy of AR/VR for language learning, it does not isolate which specific features, such as gesture-based interaction, real-time feedback, or gaze tracking. This lack of specificity constrains theory-driven, focused instructional interventions. Although AR/VR applications in education show positive effects, existing research does not specify which features, such as gesture-based interaction, real-time feedback, or gaze tracking, are responsible for the observed progress. (e.g., gesture-based interaction, real-time feedback, gaze tracking). In the absence of a rigorous model, it is challenging to distil what makes immersive learning effective and to develop focused interventions that work.

This paper presents a data-driven model that maps various immersive learning characteristics, environment type (AR, VR, Hybrid), and device modality to measurable outcomes in language skill development. The model is based on multimodal analytics and uses a massive dataset of 250K+ AR/VR language-learning sessions available on Kaggle ("English Proficiency Assessment in AR/VR"). Aligned with the Common European Framework of Reference (CEFR) framework, we developed models using gesture count, response latency, gaze duration, and interaction depth as potential mediators of learning outcomes across eight skills: reading, listening, speaking, pronunciation, vocabulary, writing, grammar, and translation. The research is more conceptual and focuses on theoretical underpinnings and design considerations, but also indicates a prospective empirical test through statistical modeling, causal model analysis, and longitudinal validation in future work.

A. Research Questions

(1) What impact does immersion (AR, VR & Hybrid) have on learners' competence in CEFR-mapped language skills: reading, listening, speaking, pronunciation, vocabulary, writing, grammar, and translation?
(2) To what extent do multimodal behavioral cues, such as gesture count, gaze duration, response latency, and interaction depth, mediate learning in AR/VR?
(3) How can AR/VR-based environments be used to support the development of equitable and adaptive language learning systems with multimodal analytics?

B. Objectives of the Study

(1) To diagnose English language proficiency scores in

doi: 10.18178/ijiet.2026.16.7.2644

various skill types from VR session learners' performance data.

(2) To investigate the influence of multimodal behavioral features (number of gestures, reaction time, and interaction depth) on English level achievement in VR environments.

(3) To study the effects of users' demographic and context factors on learning achievement and proficiency in using VR-based English teaching systems.

This study examines the burgeoning body of SLA research on learning through immersion by reporting empirical results at scale on increases in English proficiency in AR/VR contexts. Drawing on a collection of 176,000+ learner interactions, this evidence shows that VR-based assessment can more authentically assess L2 performance than traditional methods. The results demonstrate the benefits of embedding behavioral-analytic metrics, e.g., gestures, latency, and interaction depth, into language evaluation frameworks, which is crucial for developing adaptive, data-driven learning systems in global English education.

This paper aims to investigate English proficiency testing in AR/VR educational contexts with large-scale performance data. It seeks to determine how immersive interaction metrics, including reading, speaking, pronunciation, and behavioral engagement, are correlated with overall proficiency. Besides, it aims to present an SCBI-based VR arrangement and an evaluation model that combines behavioral measurement with VR technology for language tests that are far more authentic and adaptive.

II. Literature Review

A. AR/VR in Language Learning: Recent Benefits and Trends

In the last five years, AR and VR have emerged as promising tools for language education, with studies concerning both technologies growing at an unprecedented pace [6]. Importantly, immersion can give learners abundant practice in context, so learners are less anxious to use the target language [7]. Likewise, a 2022 meta-analysis provided substantial average evidence of the effect size for Extended Reality (XR) (including AR and VR) on language exposure (mean effect size ≈ 0.825), indicating extremely significant results across all levels of implementation [8].

Prior work has investigated applications of AR as opposed to VR for language learning. AR has mainly been used for vocabulary learning (often in mobile apps or with handheld devices), while VR, notably through head-mounted displays, has been employed to enable immersive speaking and listening practice [6]. The application of AR has been on the rise in language learning, traditionally used for low-level cognitive and language skills such as word identification, meaning comprehension, and spelling, while less use is observed for higher-order critical skills such as reading and writing [9]. The literature reveals motivational advantages of AR/VR in language learning, but outcomes vary depending on the technology and proficiency level. AR supports vocabulary acquisition through visual augmentation, while VR enhances speaking and listening practice. No single technology is universally superior; effectiveness depends on the targeted language competencies. However, short-term studies do not reflect long-term comprehension, and large-scale investigations are required.

B. Immersive Systems and CEFR-aligned Skill Assessment

Immersive technologies, such as augmented reality and virtual reality, are increasingly being incorporated into language learning, with particular attention to appropriate skill assessment aligned with the CEFR. This harmonization is to enable measurement, standardization, and an internationally accepted assessment of language gains in immersion settings.

Recent research suggests that VR environments can contribute to observable gains in targeted language skills, including fluency, accuracy, pronunciation, and communicative competence, as measured by CEFR-aligned pre- and post-assessments. For instance, VR interventions have generated sizable increases in learners' CEFR levels on standardized speaking tests [10].

Immersion contexts allow learners to participate in authentic communication (e.g., role plays, simulations) that reflect 'can do' descriptors within the CEFR, supporting assessment of language use within real-life situations [11]. Advances in Artificial Intelligence (AI) and ASR make testing more efficient and scalable, with real-time, CEFR-aligned feedback the norm and adaptive testing [12].

The integration of CEFR frameworks within immersive learning enables standardized, competency-based assessment in virtual environments. Evidence suggests that VR and AR effectively simulate real-world language use aligned with CEFR 'can do' descriptors, particularly for speaking and listening. But evaluating higher-order skills such as writing and mediation can be difficult. The ecosystem, however, doesn't have a standard for validating automated assessments against human-rated proficiency, even though online AI tools offer instant feedback and dynamic testing. Therefore, although CEFR-linked immersion testing is a useful development, it requires greater methodological rigor and the broader validation that comes from use in diverse settings.

C. Multimodal Interaction and Learning Analytics in Immersive Learning

Online language learning environments offer rich multimodal forms of interaction beyond plain text or audio, and researchers are investigating such forms in language learning as well. Non-verbal communication is one of the keys. During face-to-face conversation, interlocutors use a rich array of Nonverbal Cues (NVCs), such as facial displays, eye gaze, manual gestures, body position, and interpersonal distance to transmit information and regulate turn-taking, while also producing feedback [13].

Many contemporary AR/VR and XR devices can capture speech, gestures, and gaze simultaneously, thereby enabling more naturalistic interaction for language learners. Recent investigations have also confirmed that multimodal applications attract users and that system integration leads to more efficient and user-friendly interactions [14].

Multimodal Learning Analytics (MMLA) analyzes, records, and processes data streams audio, video, and gaze to monitor learning indicators such as speech rate, pronunciation, focus, and engagement during interactive tasks [15]. Combining and synchronizing multiple data streams, protecting privacy, and scaling such systems to actual classrooms is challenging [16].

Multimodal analytics, including gesture, gaze, speech, and physiological data, offer rich insights into learner engagement within immersive environments by revealing interaction patterns alongside cognitive-affective states. Most researchers view these metrics as descriptive rather than instrumental to the models with which pieces of music are learned. Note that these systems face issues of synchronization, privacy, and scalability. The field needs to further clarify how, cross-modal behaviors map onto language acquisition to inform pedagogy.

D. Equity and Accessibility in Immersive Language Learning

The literature also points to equity and accessibility concerns regarding the use of AR/VR for language learning, especially in low-resource schools and developing-world contexts. The digital divide needs to be tackled for immersive learning to become mainstream [17].

Unequal access to VR is partly attributable to disparities in learners' technological familiarity, access, and opportunities for participation. This can cause some populations to perform or engage less. And early VR language programs often attract small, self-selecting student communities, which can exacerbate educational disparities. To achieve affordability and scalability, low-cost VR solutions are investigated, but require improvements [17].

High costs of immersive technology and unequal access to reliable infrastructure are a result of digital disparity, particularly for students in low-resource and developing contexts [18]. Many immersive language tools are not universally accessible to learners with disabilities (e.g., visual, auditory, cognitive) or lack built-in accommodations for screen readers, captions, or alternative input devices [19].

However, a lack of teacher training in both technology and inclusive pedagogy can serve as a barrier to successful, equitable enactment. Economic hardship, as well as language and cultural differences, may contribute to limited access to or engagement in immersive language learning [20].

It's all about equity and access in immersive language learning, which are both critical and neglected. While barriers, including cost, digital literacy, and disability access, are recognised, there are few examples of tested solutions for inclusivity. High-cost, next-generation VR does not reconcile with the requirement for low-cost solutions in marginal areas. Additionally, content design often tends to overlook cultural and linguistic diversity. To realize more equitable approaches to language education, future work should focus on developing and evaluating accessible, culturally responsive immersive learning models.

E. Adaptive and AI-Driven Immersive Learning

An AI-adaptive, immersive learning experience is turning the tide of English language teaching through personalized, interactive, and engaging experiences. These tools use artificial intelligence, VR, and machine learning to personalize instruction, feedback, and content to individual learners' skills and needs, improving their language skills and motivation.

Using AI to analyze learner performance in real time, adaptive systems provide personalized feedback and automatically adjust content to target areas such as pronunciation, grammar, vocabulary acquisition, and fluency. Research demonstrates that students using AI-driven platforms make substantial gains in vocabulary, grammar, and conversational ability, with progress up to 35% greater than with traditional methods [21].

VR and AI-based immersive environments facilitate interactive experiences of real-world language use, thereby increasing communicative proficiency and engagement. Findings from the experimental studies indicate that in immersive AI-supported VR environments, learners demonstrate better English proficiency and motivation compared with traditional classroom settings [22].

Data privacy, algorithmic bias, and fair access are all key issues as AI is increasingly integrated in language learning [23]. The successful application is based on teacher training and meaningful connection to pedagogical aims [24].

The current AI-based technology-enhanced English teaching system that combines virtual and augmented reality has improved students' writing and oral fluency while ensuring high efficiency and stability. An AI-assisted adaptive multilingual learning system tailored to students' proficiency level, linguistic background, and learning style aids students in gaining fluency.

Personalised and AI-powered gamification. Immersive learning has been shown to significantly impact the adoption of English, thanks to tailored instruction, high-quality engagement, and meaningful practice opportunities [25]. Despite evidence of increased proficiency and motivation, continued care is needed to ensure ethical, practical, and pedagogical integration and maximize benefit [26].

AI-powered adaptive systems enhance the immersive language learning experience with instantaneous feedback and content tailored to individual proficiency levels, bolstering vocabulary retention, grammatical accuracy, and fluency in speaking. Issues at stake are relational algorithmic efficiencies, pedagogic transparency, and learner agency, as well as algorithmic bias and privacy issues. For optimal effectiveness, we recommend integrating ethical AI design with teacher oversight and strategies that account for students' needs.

F. Immersive Learning and Multimodal Interaction Theories

The use of technology to facilitate learning, known as immersive learning, employs VR, AR, and MR to foster interactive and of educational experiences. The theories and designs underlying these environments increasingly incorporate, multimodal interaction, the combination of visual, auditory, haptic, and other sensory channels of information, which facilitates learning and engagement.

Cognitive Theory of Multimedia Learning (CTML) suggests that individuals learn more when information is presented through both verbal and visual channels, but each channel has limited capacity. Instructional design should focus on reducing extraneous cognitive load and facilitating active, generative learning by attending to how learners can be involved in selecting, organizing, and integrating information [27].

Theories such as the Embodied, Enacted, Embedded, and Extended (4Es) cognition illustrate that learning in immersive environments is closely tied to physical engagement, presence, and the integration of perception and action within digital spaces [28].

Immersive worlds frequently draw on theories of constructivist, experiential, and collaborative learning, focusing on activity-based interaction, contextualisation within real-world experience, and social engagement [28]. MMLA leverages multimodal data (e.g., gameplay, eye tracking, facial expressions, physiology) to model cognitive, affective, and social learning processes. Theories such as embodied cognition and cognitive load theory are often employed to interpret this data and guide adaptive support [29].

Combining digital, physical, physiological, and psychometric data enables more precise predictions of learning outcomes and supports real-time personalized feedback [30].

Immersive learning environments are most effective when grounded in strong educational theories, particularly those related to multimodal processing, embodiment, and active social learning. New research highlighting the importance of multimodal data and analytics to personalize and optimize learning, together with the need for theory-driven design, practical frameworks to inform implementation & evaluation.

The literature reviewed shows that engaging in immersive VR spaces has the potential to revolutionise learning within the English as a Foreign Language (EFL) classroom by offering authentic, interactive, and multimodal learning environments. Previous work has shown strong gains in speaking fluency, vocabulary recall, and learner motivation, but there is fragmented evidence of how specific behavioral indicators (e.g., gesture use, gaze patterns, and latency) mediate these improvements. In addition, only a few researchers have ever investigated the combined impact of both demographic and contextual factors in VR learning systems. Hence, the present study aims to fill these gaps by exploring how VR-enacted interactions specifically contribute towards CEFR-aligned learner outcomes in English Language proficiency, identifying the underlying behavioural processes behind successful learning, and discovering an adaptive framework for more equitable and data-informed instructional support. In light of this purpose, the current study contributes to both empirical and pedagogical development in immersive language teaching.

III. Materials and Methods

A. Dataset Description

This study employed the English Proficiency Assessment in AR/VR dataset, which is publicly accessible on Kaggle (Dataset Engineer, 2024). The dataset comprises 252,782 anonymized learner-session instances collected from a a large-scale VR-based English-language learning platform. Each instance is a user session and contains demographic, behavioral, and learning-performance-related features. The main outcome measures are related to the CEFR and span eight skills: reading accuracy, listening comprehension, speaking fluency and pronunciation accuracy; vocabulary use; writing fluency and quality (including written range); grammatical accuracy; translation ability.

B. Data Preparation

The original dataset contained 252,782 learner sessions. After a multi-step cleaning process to ensure data integrity and analytical quality, a final sample of 176,911 sessions was retained. Incomplete records sessions missing any of the eight CEFR-aligned proficiency scores or key behavioral variables (gesture count, response latency, interaction depth, session duration) were excluded ($n = 48{,}221$). Second, outlier sessions with extreme or technically implausible values were excluded: session duration from less than 2 min to more than 120 min ($n = 12{,}500$), response time from below 100 ms to above 5000 ms ($n = 9150$), and number of gestures in the top s percentile ($n = 6000$). Lastly, sessions that showed bot-like, duplicated, or similar interaction behavior were rejected ($n = 5000$). This cleaning approach yielded a robust, high-quality dataset suitable for reliable statistical analysis while preserving ecological validity.

C. Variables and Measures

Dependent Variables (Language Proficiency Scores): CEFR-aligned skill scores measuring proficiency in reading, listening, speaking, pronunciation, vocabulary, writing, grammar, and translation. Independent Variables (VR Behavioral Indicators): Gesture count, response latency, average interaction depth, and session duration, which reflect the learner's level of engagement and interaction in the immersive VR environment. Covariates (Demographic and Contextual): Age group, gender, native language, education level, country/region, and device type (VR headset, tablet, or desktop).

Embodied Cognition, Cognitive Load Theory, and Multimodal Learning Analytics are key theoretical frameworks for designing immersive language environments. Nevertheless, research tends to employ these theories post hoc rather than experimental investigations based on hypotheses. Only a small number of studies have investigated whether decreased cognitive demand in VR is associated with language gains or how gestures may support syntactic processing. Further studies are suggested to adopt a theory-based experimental design, compare two or more training programs, and evaluate the effectiveness of learning and cognitive/affective processing.

D. Research Design

The study employed a quantitative correlational research design to examine the association between learners' behavioral engagement and their English proficiency outcomes when using VR for learning. Comparison between device types and demographic groups was used to show variation in performance and engagement. The objective of the study was to identify patterns of proficiency gains related to general interaction features in VR/AR.

E. Population and Context of Study

The research used data from a large international sample of 176,911 students participating in AR/VR-supported English language learning programs worldwide. The English Proficiency Assessment in AR/VR dataset (Kaggle, 2024) presents a heterogeneous group with immersive-based learners from several regions. While country identifiers were de-identified, the overall speakers represented a diverse learner population, supporting a sample representative of worldwide ELLs. Learners' proficiency levels ranged from basic to advanced, allowing a wide spectrum of language skill development in reading, listening, speaking, and pronunciation to be examined.

IV. Result and Discussion

In this section, we present the results of the statistical analysis and discuss their implications for immersive AR/VR language learning. Findings are structured from descriptive summaries to inferential models, then to an integrated discussion that relates quantitative findings to the theoretical framework.

Learners' performance was summarized through descriptive analyses on core English proficiency domains: reading, listening, speaking, pronunciation, vocabulary, writing, grammar, and translation. Behavioral factors included the number of gestures, response latency, depth of interaction within sessions, and session duration.

Tables 1 and 2 show that they are balanced, as the mean scores for the reading and listening approaches are about 70%, indicating moderate competence among VR learners. The low standard deviations indicate the consistency of learners' results. Median values of 70 and upper quartiles in the 80–85 range illustrate a good understanding among high performers, while lower {minimum} values indicate more help is required for weaker students. Taken together, these findings suggest that VR tools offer a standardized setting for measuring receptive language skills, promoting equitable skill development among the participants.

Table 1. Descriptive statistics—core proficiency variables (N = 176,911)

Variable	N	Mean	SD	Minimum
Reading Accuracy	176,911	70.01	17.30	40.00
Listening Comprehension	176,911	69.90	14.71	20.00

Table 2. Descriptive statistics—percentile metrics (N = 176,911)

Variable	25th Percentile	Median	75th Percentile	Maximum
Reading Accuracy	55.10	70.00	85.00	100.00
Listening Comprehension	59.90	70.00	80.20	100.00

Table 3. Descriptive statistics for oral proficiency, behavioral metrics, and session duration (N = 176,911)

Variable		Mean	SD	Minimum	25th Percentile	Median
oral and pronunciation proficiency	Speaking Fluency Score	0.40	0.20	0.00	0.24	0.39
	Pronunciation Score	0.60	0.20	0.01	0.46	0.61
oral proficiency and pronunciation	Speaking Fluency Score	0.40	0.20	0.00	0.24	0.39
	Pronunciation Score	0.60	0.20	0.01	0.46	0.61
gesture and response metrics	Gesture Count	29.54	17.32	0.00	15.00	29.00
	Response Latency (ms)	600.85	197.84	150.00	465.00	600.00
session duration metrics	Session Duration (min)	25.02	7.96	5.00	19.60	25.00

Table 4. Descriptive statistics for vocabulary, interaction depth, and session duration percentile metrics (N = 176,911)

Variable		75th Percentile	Maximum
vocabulary proficiency	Vocabulary Usage Score	68.50	100.00
interaction depth metrics	Gesture Count	45.00	59.00
	Response Latency (ms)	735.00	1500.00
	Average Interaction Depth	11.10	30.00
session duration percentile distribution	Session Duration (min)	30.40	60.00

Tables 3 and 4 reveal the moderate oral proficiency among learners, with speaking fluency mean score (0.40), pronunciation accuracy mean score (0.60), demonstrating emergent communicative competence within VR-based environments. The small standard deviations indicate that performance across subjects is quite consistent. Vocabulary usage scores, with a 75th percentile of 68.50 and a maximum of 100, indicate lexical diversity among advanced learners. Together, these findings indicate that immersive practice promotes pronunciation accuracy and vocabulary acquisition, underscoring the potential value of VR for active language production and sustained oral proficiency development.

Writing, grammar, and translation scores are reported as percentages (0–100). The results indicate moderate writing performance (M = 60.05, SD = 14.90) and grammar accuracy (M = 62.49, SD = 18.73), with higher variability in translation skills (M = 50.01, SD = 23.99). The interquartile ranges suggest a steady progression toward higher linguistic accuracy and translation proficiency among VR learners.

Behavioral metrics represent learners' engagement and interactivity within the VR environment. The mean gesture count (M = 29.54, SD = 17.32) and moderate response latency (M = 600.85 ms, SD = 197.84) indicate active yet varied learner participation. Average interaction depth (M = 7.90, SD = 7.22) shows consistent engagement levels with considerable variability, reflecting different exploration patterns across VR sessions.

Session duration represents the total time learners spent actively engaged in each VR-based English learning session. The mean session length (M = 25.02 min, SD = 7.96) suggests consistent engagement across sessions, with most learners spending approximately 20–30 min per session. The maximum recorded duration of 60 min indicates sustained attention in extended learning activities for a subset of users.

The correlation matrix revealed no substantive correlations between proficiency and behavioral measures (all $r < 0.01$). This suggests that language skill performance in VR is not dependent on raw quantitative behavioral measures such as gesture frequency, response latency, or interaction depth. In other words, students' progress is not directly linked to how much or quickly they move in VR, suggesting that qualitative features of our activities (e.g., task design, instructional setting) might be more important than raw behavioural engagement.

To examine whether the type of device used by learners influenced language proficiency outcomes, a Multivariate Analysis of Variance (MANOVA) was conducted. The eight CEFR-aligned proficiency scores (reading, listening, speaking, pronunciation, vocabulary, writing, grammar, and translation) served as dependent variables. At the same time, device type (VR headset, tablet, AR mobile, desktop) was entered as a fixed factor. The results, presented in Table 5, indicate no statistically significant multivariate effect of device type on language proficiency outcomes.

Table 5. Multivariate Analysis of Variance (MANOVA) results for the effect of device type on language proficiency outcomes (N = 176,911)

Effect	Test Statistic	Value	Num DF	Den DF	F Value	Sig. (p)
Intercept	Wilks' Lambda	0.0352	8	176,900	605,572.35	/
	Pillai's Trace	0.9648	8	176,900	605,572.35	/
	Hotelling-Lawley Trace	27.3860	8	176,900	605,572.35	/
	Roy's Greatest Root	27.3860	8	176,900	605,572.35	/
Device Type	Wilks' Lambda	0.9999	24	513,064.42	0.85	0.6777
	Pillai's Trace	0.0001	24	530,706.00	0.85	0.6777
	Hotelling-Lawley Trace	0.0001	24	383,276.14	0.85	0.6777
	Roy's Greatest Root	0.0001	8	176,902.00	1.96	0.0471

To investigate the extent to which behavioural engagement metrics predict overall English proficiency in VR-based language learning, a multiple linear regression analysis was conducted. The behavioural predictors included gesture count, response latency, average interaction depth, and session duration. Overall proficiency, calculated as the mean of the eight CEFR-aligned skill scores, served as the dependent variable. The regression coefficients, standard errors, and significance levels are presented in Table 6.

Multiple regression analysis explored predictors of mean proficiency scores. Results indicated that session duration was a statistically significant but weak predictor ($p < 0.05$), while gesture count and response latency were not significant.

To assess the precision of the regression estimates, 95% confidence intervals were calculated for each predictor. As shown in Table 7, the confidence intervals for gesture count, response latency, and average interaction depth include zero, indicating non-significant effects on proficiency. Only session duration demonstrates a positive but minimal effect, with its lower bound approaching zero.

Mediation analysis revealed no significant indirect effect of response latency on the relationship between gesture count and proficiency.

Table 6. Regression coefficients and significance (N = 176,911)

Predictor	Coefficient (B)	Std. Error	t-value	Sig. (p)
Constant	45.9497	0.066	693.558	0.000
Gesture Count	0.0004	0.001	0.458	0.647
Response Latency	0.0000116	0.000068	0.170	0.865
Avg. Interaction Depth	0.0006	0.002	0.316	0.752
Session Duration	0.0035	0.002	2.047	0.041

Table 7. Confidence intervals for regression coefficients (N = 176,911)

Predictor	95% CI Lower	95% CI Upper
Constant	45.820	46.080
Gesture Count	−0.001	0.002
Response Latency	−0.000	0.000
Avg. Interaction Depth	−0.003	0.004
Session Duration	0.000	0.007

Model Summary-R^2 = 0.000, Adjusted R^2 = 0.000, $F(4, 176906) = 1.13$, $p = 0.340$

The regression model examined the effect of four behavioral metrics, gesture count, response latency, average interaction depth, and session duration, on overall English proficiency. The predictors collectively explained negligible variance ($R^2 = 0.000$, $p > 0.05$). Only session duration showed a weak positive effect ($B = 0.0035$, $p = 0.041$), indicating that longer engagement slightly improves proficiency.

The regression model yielded an R^2 of 0.000, indicating that behavioral metrics collectively explain virtually no variance in proficiency scores. This suggests that VR language learning is not merely a function of interaction frequency or speed but is likely influenced by qualitative factors such as task design, contextual relevance, feedback quality, and learner agency.

Our findings challenge simplistic interpretations of embodied cognition, suggesting that physical interaction alone is insufficient for linguistic gains without meaningful engagement. It indicates that the VR courseware should integrate physical actions with communication acts to improve language processing. Future systems should focus more on the relevance of motion to language use than on its magnitude.

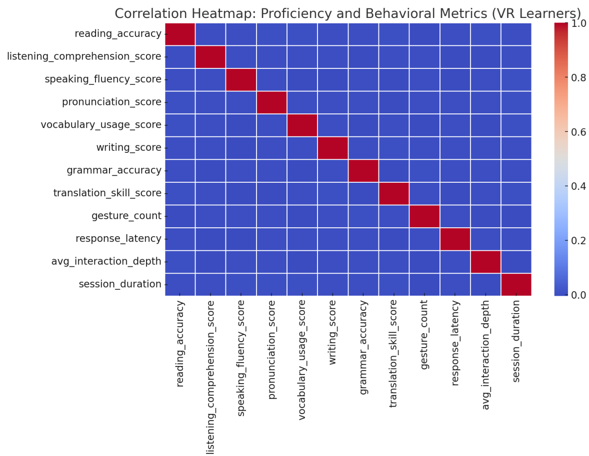

Fig. 1. Correlation heatmap of proficiency and behavioral metrics (VR learners).

Fig. 1 illustrates the correlation heatmap between English language proficiency measures and behavioral engagement variables among VR learners. The color gradient represents the Pearson correlation coefficients, with dark red indicating stronger positive relationships and dark blue representing weak or no correlation.

The heatmap shows minimal inter-variable relationships across all proficiency and behavioral metrics. The diagonal dominance of red squares indicates perfect self-correlations, while the surrounding blue regions signify near-zero relationships between distinct variables. This confirms that no single behavioral feature (gesture count, latency, interaction depth, or session duration) strongly correlates with proficiency outcomes such as reading, listening, or writing accuracy.

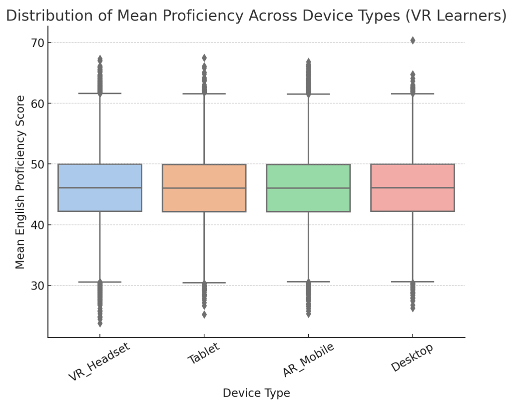

Fig. 2. Distribution of mean proficiency across device types (VR learners).

Fig. 2 shows the distribution of mean English proficiency scores across four device types (VR Headset, Tablet, AR Mobile, and Desktop). Each boxplot represents the median, interquartile range, and outliers for proficiency performance among VR learners using different hardware setups.

The boxplot indicates that mean proficiency scores remain consistent across all device types, with median values clustering around 45–50. The small interquartile ranges and overlapping distributions demonstrate minimal variance, aligning with the MANOVA findings that device type had no statistically significant effect on proficiency ($p > 0.05$). This consistency highlights that learning outcomes in immersive environments are hardware-independent, confirming that VR-based English learning can be effectively delivered across multiple devices without compromising instructional quality or learner performance.

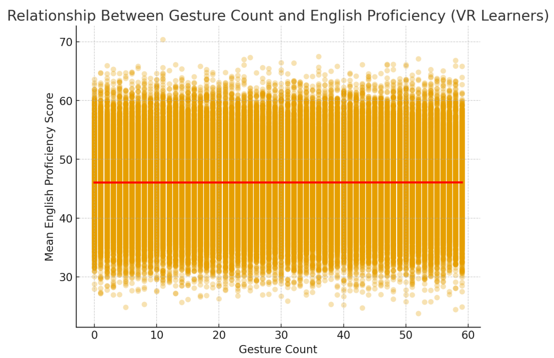

Fig. 3. Relationship between gesture count and English proficiency (VR learners).

Fig. 3 depicts a scatter plot of the relationship between gesture count and mean English proficiency scores among VR learners. The red regression line indicates the line of best fit. Data points represent individual learners, with gesture frequency on the x-axis and proficiency score on the y-axis. The scatterplot shows a nearly flat regression line, indicating an extremely weak relationship between gesture count and English proficiency ($B = 0.0004$, $p = 0.647$). The distribution of data points is wide but uniform, suggesting that increased physical interaction (gestures) does not correspond to measurable improvements in proficiency performance. This finding aligns with the regression results, confirming that gesture-based engagement alone does not predict language acquisition in VR contexts. These results highlight that qualitative engagement, such as contextual feedback or immersive task design, may be more influential than raw motion frequency for skill development in virtual learning environments.

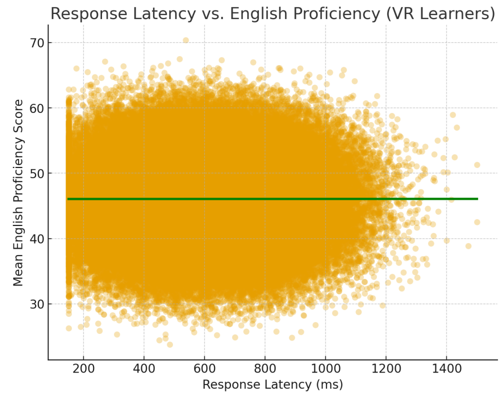

Fig. 4. Response latency vs. English proficiency (VR learners).

Fig. 4 presents a scatter plot illustrating the relationship between response latency and mean English proficiency among VR learners. Each data point represents a learner, with response latency (in milliseconds) plotted on the x-axis and mean proficiency on the y-axis. The green regression line represents the best-fit trend across all observations. The scatterplot demonstrates a near-zero slope, revealing no meaningful relationship between response latency and proficiency ($B = 1.15\text{e}{-05}$, $p = 0.865$). Learners with shorter or longer latencies show comparable proficiency levels, suggesting that response speed does not contribute significantly to language outcomes in VR learning environments. This finding reinforces the regression analysis, indicating that behavioral timing metrics, such as latency, are not predictive of English proficiency. The wide, uniform spread of data implies that learners' performance remains stable regardless of their reaction time during interactive VR tasks, highlighting the independence of linguistic skill acquisition from temporal engagement metrics.

Overall, our results demonstrate the potential of immersive AR/VR systems for ensuring a reliable, consistent, and equitable environment for language skill assessment. Although the use of behavioural indexes extends the conception of learner engagement, linguistic attainment is still more directly predicted by cognitive and affective aspects of learning. The findings are consistent with developments concerning immersive learning theory and multimodal learning analytics, and suggest that integrated AR/VR environments can serve as a valuable adjunct to traditional CEFR-based assessment systems.

V. Conclusion

In this study, we proposed and empirically investigated a

VR-based English assessment model that combines multimodal behavioral data with CEFR-aligned language performance measures. Based on a large dataset of 176,911 learner interactions, results indicate moderate proficiency levels across reading, listening, writing, and grammar tasks; however, these outcomes were not predictive of higher engagement in subsequent tasks or assessments.

The next step could also consider cross-linguistic validation by testing the model across other languages and proficiency levels to increase generalization. In the future, it remains to be seen how task design factors, e.g., narrative coherence, type of feedback, or cognitive challenge, moderate the relation between engagement and language gains. Moreover, by using Generative AI and NLP for automated speech feedback and personalised learning pathways, it opens up the potential of turning immersive assessment into an intelligent system, one that learns continuously. Interdisciplinary efforts involving linguists, cognitive scientists, and technologists are needed to create ethically responsible, inclusive data-driven AR/VR learning environments that offer new standards or understanding of how language proficiency is examined and learned in virtual settings.

The behavioral measurements, such as gesture count and response latency, explained a negligible amount of the variance in proficiency scores ($R^2 = 0.000$), according to the regression model. This is in stark contrast to the belief that more time and active involvement necessarily result in better learning. This is consistent with cognitive load theory and can be taken to mean that ongoing bodily interactions among children may subvert verbal processing, provided the task or tasks are not well-designed. It is also consistent with embodied cognition, predicting that gestures facilitate learning, but only when in harmony with goals. The findings underscore that the potency of behavioral engagement is contingent upon pedagogical integration, a premise that has scarcely been examined empirically in prior work.

This study demonstrates that VR-mediated language learning can engender sustainable skills through the quality of immersive tasks, rather than on mere behavioural engagement. The gains through learning will be the pedagogical quality and communicative task-based validity, with special emphasis on the distinction between activity and real practice in multimodal language study. A personal adaptive, contextualised scenario-based model for language learning: Developing technology for language teaching and learning.

VI. Limitations of the Study

The qualitative aspects of the study, including both positive and negative attitudes, depend on the quality of the learning activities. While this study was based on a large dataset and offered a novel framework for VR-mediated ELP (English Language Proficiency) assessment, it had some limitations. First, the study used secondary data from previously conducted VR learning sessions, which limited control over instruction conditions, learner characteristics, and contextual factors such as task difficulty and motivation. However, computing total sums (e.g., total count) might obscure meaningful temporal/sequential interaction patterns that are more indicative of learning outcomes. Second, the study focused solely on quantitative performance and behaviour indicators, ignoring feedback or emotional engagement, which might also help explain differences in proficiency. Third, technical parameters, such as headset resolution and latency across devices and settings, as well as environmental parameters, were not normalized, which might have caused a slight mismatch in measurements. Lastly, the framework is pending longitudinal validation and real-world experimental demonstration, which will be necessary to establish causal links and generalization across different learning populations and linguistic environments. Notably, the use of aggregate behavioral metrics may have obscured meaningful temporal patterns in learner interaction, which could explain the weak relationship between these metrics and proficiency outcomes.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Saima Mehboob (University of Bari Aldo Moro) analyzed and designed the study, conducted the statistical analyses, and drafted the manuscript. Professor Alberto Fornasari coordinated the design of the study; inspired, supported, and revised the theoretical background; aided in data analysis, interpretation, and drafting of the final version of the manuscript. Both authors read and approved the final manuscript for publication.

Funding

This work was financially supported by the PNRR Scholarship Programme of Italy, within the National Recovery and Resilience Plan (PNRR). The authors would like to express their thanks for this support, which has facilitated the successful completion of this study.

Acknowledgment

The author would like to thank Prof. Alberto Fornasari of the University of Bari Aldo Moro for his invaluable research supervision, guidance, and support during the development of this study. This research was funded by a PNRR Scholarship under the Italian National Recovery and Resilience Plan, enabling its implementation. The author supported the English Proficiency Assessment in AR/VR Dataset (Kaggle, 2024) for empirical analysis and involved academic peers during the manuscript draft.

References

[1] Z. Sadigzade, "Immersive and gamified approaches: VR/AR in language learning," *Porta Universorum*, vol. 1, no. 6, pp. 15–24, Aug. 2025. doi: 10.69760/portuni.0106002

[2] Y.-J. Lan, "Language learning in virtual reality: Theoretical foundations and empirical practices," in *Contextual Language Learning*, Y.-J. Lan and S. Grant, Eds., Chinese Language Learning Sciences, Singapore: Springer Singapore, 2021, pp. 1–21. doi: 10.1007/978-981-16-3416-1_1

[3] A. Parmaxi and A. A. Demetriou, "Augmented reality in language learning: A state-of-the-art review of 2014–2019," *J. Comput. Assist. Learn.*, vol. 36, no. 6, pp. 861–875, Dec. 2020. doi: 10.1111/jcal.12486

[4] R. Shadiev and Q. Liang, "A review of research on AR-supported language learning," *Innov. Lang. Learn. Teach.*, vol. 18, no. 1, pp. 78–100, Jan. 2024. doi: 10.1080/17501229.2023.2229804

[5] H. Hu, G. Liu, and T. Xie, "Multimodal interaction in virtual reality supported education: A systematic review," *Interact. Learn. Environ.*, vol. 33, no. 1, pp. 170–191, Jan. 2025. doi: 10.1080/10494820.2024.2342993

[6] M. M. Zhang, H. Hashim, and M. M. Yunus, "Analyzing and comparing augmented reality and virtual reality assisted vocabulary learning: A systematic review," *Front. Virtual Real.*, vol. 6, 1522380, May 2025. doi: 10.3389/frvir.2025.1522380

[7] X. Huang, D. Zou, G. Cheng, and H. Xie, "A systematic review of AR and VR enhanced language learning," *Sustainability*, vol. 13, no. 9, 4639, Apr. 2021. doi: 10.3390/su13094639

[8] J. Chen, J. Dai, K. Zhu, and L. Xu, "Effects of extended reality on language learning: A meta-analysis," *Front. Psychol.*, vol. 13, 1016519, Sept. 2022. doi: 10.3389/fpsyg.2022.1016519

[9] H. Chang, J. Park, and J. Suh, "Virtual reality as a pedagogical tool: An experimental study of English learner in lower elementary grades," *Educ. Inf. Technol.*, vol. 29, no. 4, pp. 4809–4842, Mar. 2024. doi: 10.1007/s10639-023-11988-y

[10] M. Dooly, T. Thrasher, and R. Sadler, "'Whoa! incredible!' language learning experiences in virtual reality," *RELC J.*, vol. 54, no. 2, pp. 321–339, Aug. 2023. doi: 10.1177/00336882231167610

[11] A. Berns *et al.*, "'Let's date!' A 360-degree video application to support foreign language learning," in *Proc. CALL and complexity—Short Papers from EUROCALL 2019*, 1st ed. 2019, pp. 39–44. doi: 10.14705/rpnet.2019.38.983

[12] H. Yannakoudakis, Ø. E. Andersen, A. Geranpayeh, T. Briscoe, and D. Nicholls, "Developing an automated writing placement system for ESL learners," *Appl. Meas. Educ.*, vol. 31, no. 3, pp. 251–267, July 2018. doi: 10.1080/08957347.2018.1464447

[13] A. Gruber and R. Kaplan-Rakowski, "Verbal and nonverbal communication in high-immersion virtual reality for language learners," in *Proc. Intelligent CALL, Granular Systems and Learner Data: Short Papers from EUROCALL 2022*, 1st ed. B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouësny, and M. J. Whelpton, Eds. 2022, pp. 129–134. doi: 10.14705/rpnet.2022.61.1447

[14] Z. Wang, H. Wang, H. Yu, and F. Lu, "Interaction with gaze, gesture, and speech in a flexibly configurable augmented reality system," *IEEE Trans. Hum.-Mach. Syst.*, vol. 51, no. 5, pp. 524–534, Oct. 2021. doi: 10.1109/THMS.2021.3097973

[15] H. Ouhaichi, D. Spikol, and B. Vogel, "Learning Swedish with AI: Exploring multimodal learning analytics in spoken language acquisition," in *Proc. Methodologies and Intelligent Systems for Technology Enhanced Learning, 14th International Conference*, vol. 1171, C. Herodotou, S. Papavlasopoulou, C. Santos, M. Milrad, N. Otero, P. Vittorini, R. Gennari, T. Di Mascio, M. Temperini, and F. De La Prieta, Eds. in Lecture Notes in Networks and Systems, Cham: Springer Nature Switzerland, vol. 1171, 2024, pp. 178–189. doi: 10.1007/978-3-031-73538-7_16

[16] W. Chango, J. A. Lara, R. Cerezo, and C. Romero, "A review on data fusion in multimodal learning analytics and educational data mining," *WIREs Data Min. Knowl. Discov.*, vol. 12, no. 4, e1458, July 2022. doi: 10.1002/widm.1458

[17] A. Yudintseva, "Virtual reality affordances for oral communication in English as a second language classroom: A literature review," *Comput. Educ. X Real.*, vol. 2, 100018, 2023. doi: 10.1016/j.cexr.2023.100018

[18] Z. Sadigzade, "Immersive and gamified approaches: VR/AR in language learning," *Porta Universorum*, vol. 1, no. 6, pp. 15–24, Aug. 2025. doi: 10.69760/portuni.0106002

[19] J. Nijakowska, "Increasing accessibility in language education—the SCALED project," *Appl. Linguist. Pap.*, vol. 1/2024, no. 28, pp. 63–73, Mar. 2024. doi: 10.32612/uw.25449354.2024.1. pp.63-73

[20] J. C Meniado, "Sustaining inclusive, equitable, and accessible language education in a COVID-19 endemic world," *RELC J.*, vol. 54, no. 1, pp. 3–6, Apr. 2023. doi: 10.1177/00336882231157495

[21] M. V. Kumar, N. Prakash, P. S, M. Manchanda, and R. Changala, "Harnessing AI for English language Learning through personalized methods and adaptive technologies," in *Proc. 2025 International Conference on Intelligent Systems and Computational Networks (ICISCN)*, Bidar, India: IEEE, Jan. 2025, pp. 1–5. doi: 10.1109/ICISCN64258.2025.10934428

[22] H. Chang, J. Park, and J. Suh, "Virtual reality as a pedagogical tool: An experimental study of English learner in lower elementary grades," *Educ. Inf. Technol.*, vol. 29, no. 4, pp. 4809–4842, Mar. 2024. doi: 10.1007/s10639-023-11988-y

[23] N. Hockly, "Artificial intelligence in English language teaching: The good, the bad and the ugly," *RELC J.*, vol. 54, no. 2, pp. 445–451, Aug. 2023. doi: 10.1177/00336882231168504

[24] D. Pérez-Jorge, E. Olmos-Raya, A. I. González-Contreras, and I. Pérez-Pérez, "Technologies applied to education in the learning of English as a second language," *Front. Educ.*, vol. 10, 1481708, Mar. 2025. doi: 10.3389/feduc.2025.1481708

[25] Y. Meng, "Design of English teaching system based on artificial intelligence technology," in *Proc. the 2024 International Conference on Artificial Intelligence and Future Education*, Shanghai Guangdong China: ACM, Nov. 2024, pp. 109–113. doi: 10.1145/3708394.3708414

[26] J. C. Lawrance *et al.*, "Developing an AI-assisted multilingual adaptive learning system for personalized English language teaching," in *Proc. 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India: IEEE, Mar. 2024, pp. 428–434. doi: 10.1109/ICACCS60874.2024.10716887

[27] M. Mulders, J. Buchner, and M. Kerres, "A framework for the use of immersive virtual reality in learning environments," *Int. J. Emerg. Technol. Learn. IJET*, vol. 15, no. 24, p. 208, Dec. 2020. doi: 10.3991/ijet.v15i24.16615

[28] T. Kim, J. Planey, and R. Lindgren, "Theory-driven design in metaverse virtual reality learning environments: Two illustrative cases," *IEEE Trans. Learn. Technol.*, vol. 16, no. 6, pp. 1141–1153, Dec. 2023. doi: 10.1109/TLT.2023.3307211

[29] M. Sobocinski, D. Dever, M. Wiedbusch, F. Mubarak, R. Azevedo, and S. Järvelä, "Capturing self-regulated learning processes in virtual reality: Causal sequencing of multimodal data," *Br. J. Educ. Technol.*, vol. 55, no. 4, pp. 1486–1506, July 2024. doi: 10.1111/bjet.13393

[30] A. Emerson, E. B. Cloude, R. Azevedo, and J. Lester, "Multimodal learning analytics for game-based learning," *Br. J. Educ. Technol.*, vol. 51, no. 5, pp. 1505–1526, Sept. 2020. doi: 10.1111/bjet.12992