International Journal of
Computational Linguistics Research

DLINE JOURNALS

Print ISSN: 0976-416X
Online ISSN: 0976-4178

IJCLR 2026: 17 (4)
https://doi.org/10.6025/ijclr/2026/17/4/227-251

Multimodal Learning Analytics in English Education : A Comprehensive Profiling of Language Skills Using the AI-IOT Dataset

Maleerat Sodanil
Faculty of Information Technology, King Mongkut's University
of Technology North Bangkok, Bangkok
Thailand
maleerat.s@it.kmutnb.ac.th

ABSTRACT

Multimodal Learning Analytics (MMLA) integrated with Artificial Intelligence (AI) and the Internet of Things (IoT) offers transformative potential for understanding complex language learning processes. This study presents a comprehensive profiling of English language skills using an AI–IoT dataset comprising 12,500 learning sessions from 2,479 unique students. By employing a robust analytical framework including repeated measures ANOVA, Pearson correlation, Principal Component Analysis (PCA), and K-means clusteringwe investigated cohort level proficiency distributions, inter skill relationships, and within-student skill imbalances across six domains: Vocabulary, Grammar, Reading, Writing, Listening, and Speaking. Results indicate that while cohort level mean scores across the six skills are statistically indistinguishable, substantial within student proficiency gaps exist (mean maximum to minimum gap = 11.85 points). High positive correlations (r = 0.650–0.709) and PCA reveal a dominant general proficiency dimension accounting for 73.08% of the standardized variance. Furthermore, unsupervised clustering primarily separates learners by overall performance levels rather than distinct skill-specialization patterns, and multivariate analyses (MANOVA) show no significant omnibus profile differences across gender, grade level, or learning mode. These findings underscore the limitation of relying solely on aggregate averages and highlight the critical need for individualized, multidimensional learner profiling to inform AI–IoT-enabled personalized English education.

Keywords: Multimodal Learning Analytics (MMLA), AI–IoT Integration, English Language Proficiency, Learner Profiling, Educational Data Mining, Personalized Learning

Received: 25 March 2026, Revised: 10 June 2026, Accepted: 22 June 2026

Copyright: DLINE

1. Introduction

Online learning analytics, also described as educational data mining, involves the systematic collection, analysis, and interpretation of data generated through digital learning activities. Such data can include learners' interactions with course materials, participation in discussions, assignment completion, and assessment performance [1, 2]. By examining these signals, learning analytics can help characterize learner engagement

and performance and provide evidence to support educational decision making.

The development of artificial intelligence (AI), the Internet of Things (IoT), and multimedia technologies has expanded the scope of learning analytics beyond single source data. In English education, these developments create opportunities to combine learner behaviour information with multimodal evidence to better understand learning processes and outcomes. This broader approach provides the conceptual basis for using AI- and IoT-enabled data to profile language learning.

2. Earlier Studies

2.1 Multimodal Learning Analytics in Education

Multimodal learning analytics (MMLA) is an emerging area that seeks to provide a more comprehensive account of learning by integrating different forms of data. In contrast to unimodal approaches, which typically capture only one dimension of learner activity, MMLA can combine signals that reflect both observable behaviour and cognitive affective states [3]. For example, visual, auditory, and physiological signals may offer complementary evidence about engagement, cognitive load, and collaboration, thereby supporting more informed assessment and personalized interventions [4, 5].

The value of multimodal analysis also extends to complex learning situations. By modelling group dynamics and emotional responses, MMLA can support collaborative learning and feedback systems [6, 7]. Integrating diverse signals may further reveal context-specific learning patterns that are difficult to detect using a single modality, improving the interpretability and practical value of educational analytics [8, 9, 10]. A systematic exploration of technology enhanced MMLA in higher education similarly reflects the growing research interest in this area [11].

2.2 AI–IoT Integration for Adaptive English Learning

AI and IoT provide a technological foundation for collecting and interpreting multiple forms of learning data. Wang, X [12] proposes an AI-driven learning behaviour analysis and modelling framework intended to support adaptive learning in English education. The framework combines machine learning models with IoT technology and multimedia capabilities to examine students' online behavioural performance during learning and evaluate their learning outcomes [12]. This approach connects the observation of learning activity with the analysis of outcomes, with the broader aim of supporting adaptation to learner needs.

More generally, the integration of AI and IoT enables intelligent systems to combine information from different sources to inform decision making. In educational settings, including those associated with smart cities, these technologies may help monitor learner engagement and inform efforts to optimize learning experiences [13] Their usefulness, however, depends on the quality and relevance of the data collected and on how appropriately the resulting analyses are interpreted.

2.3 Multimodal Semantic Understanding and Language Technologies

Multimodal semantic understanding extends text-only approaches by integrating image, text, and audio information. This integration can improve information completeness and accommodate heterogeneous inputs, while supporting contextualized and interactive learning scenarios [14]. It also offers a technical pathway for strengthening language transferability and semantic depth [15]. These capabilities are relevant to language education because meaning may be conveyed through written language, speech, and visual context.

Recent developments in image text joint learning, including VisualBERT [16] and speech text alignment, including wav2vec [17], have broadened the application of natural language processing to multimodal tasks [18]. Such methods aim to extract and align information across modalities to improve semantic understanding [19]. For example, VisualBERT uses a pretrained convolutional neural network (CNN) to represent image information and a BERT encoder for textual inputs; the resulting representations are brought into a shared semantic space [20, 21]. This example illustrates how multimodal representations can connect visual context and language within a common framework.

2.4 Multimodal Data Fusion and Personalized Learning

Data fusion is an important component of intelligent educational systems because it brings together complementary information from different modalities. Ji, X [22] examined multimodal data fusion approaches in intelligent education by considering the research landscape, data types, and relevant fusion strategies. The stated objective of this line of work is to develop personalized services that respond to individual learners' needs [22]. In principle, the integration of learner interaction data with other relevant signals can provide a more differentiated profile than any single source alone.

2.5 Synthesis and Implications for English Education

Taken together, the literature presented here positions MMLA as a means of linking learning activity, multimodal semantic information, and learner outcomes. AI and IoT can support the acquisition and analysis of learning-related data, while multimodal models can integrate linguistic, visual, and auditory information. Data fusion, in turn, provides a basis for learner profiling and personalized support. For English education, this combination offers a conceptual foundation for examining language learning processes through multiple complementary signals rather than relying exclusively on conventional text-based or single modality indicators.

The cited discussion establishes the relevance of multimodal and AI–IoT approaches, but it does not specify the structure, variables, sample, or experimental results of the AI–IoT dataset referenced in the title. Accordingly, dataset specific claims and empirical conclusions should be added only after the dataset and analysis procedures are documented. A complete journal manuscript would also benefit from explicit research questions, a detailed dataset description, the proposed analytical workflow, evaluation criteria, and a results section grounded in observed evidence.

2.6 Research Issues and Research Questions

2.61 Research issues

The integration of artificial intelligence (AI), the Internet of Things (IoT), and multimodal learning analytics (MMLA) provides opportunities to develop more comprehensive representations of English language learning. However, several research issues remain important when translating these technologies into reliable learner-profiling systems.

Research Issue 1: Integrating heterogeneous learning information. Traditional educational analytics often rely on individual assessment scores or learning platform interactions. MMLA offers a broader perspective by integrating behavioral, attentional, engagement related, and academic information. The challenge is to determine how these different indicators can be combined meaningfully without introducing redundant features or relying on unvalidated measurements.

Research Issue 2: Identifying individual differences beyond average performance. The reported mean scores are closely aligned across Vocabulary, Grammar, Reading, Writing, Listening, and Speaking. Nevertheless, the mean within-student maximum-to-minimum gap is 11.85 points, with a median of 10.94 points. This indicates that similar cohort averages can coexist with substantial differences in individual skill profiles. A central research issue is therefore how to identify these differences and translate them into interpretable learner profiles.

Research Issue 3: Distinguishing general proficiency from skill-specific characteristics. The six language skills exhibit positive correlations ranging from 0.650 to 0.709. Furthermore, the first principal component accounts for 73.08% of standardized variance. These findings suggest a strong shared proficiency dimension, although the remaining variation indicates that domain-specific information may still be important. The analytical challenge is to balance overall proficiency estimation with the identification of individual skill strengths and weaknesses.

Research Issue 4: Establishing reliable contextual associations. Learner characteristics, learning activities, attention, engagement, motivation, stress, learning mode, and device type may provide useful contextual information. However, observational associations do not establish causal relationships. Robust statistical

procedures and transparent reporting are necessary to determine which variables are associated with particular language skills.

Research Issue 5: Validating learner-profile clusters. The reported K-means solution selects two clusters with a silhouette score of 0.353. Their nearly parallel skill profiles suggest that the clusters primarily distinguish overall score levels rather than fundamentally different skill-specialization patterns. Further stability analysis is required before these clusters can inform instructional decisions.

Research Issue 6: Ensuring reproducibility and measurement validity. The manuscript reports no missing cells or exact duplicate rows, but data completeness alone does not establish assessment reliability, construct validity, or comparability across learner groups. In addition, the source contains a discrepancy between 34 features and 35 variables. This must be resolved before final publication.

2.62 Research questions

The following questions provide a coherent framework for the study.

• RQ1: How do students' proficiency levels differ across Vocabulary, Grammar, Reading, Writing, Listening, and Speaking?

• RQ2: What relationships exist among the six English-language skills, and to what extent can their shared variation be represented by a common proficiency dimension?

• RQ3: What degree of within-student skill imbalance exists, and which language domains most frequently emerge as relative strengths and weaknesses?

• RQ4: Do multidimensional English-language proficiency profiles differ across gender, grade level, and learning mode?

• RQ5: Can unsupervised clustering identify interpretable and stable learner profiles based on the six language skills?

• RQ6: What adjusted associations exist between learning-context variables, behavioral indicators, and individual English-language skills?

• RQ7: How can validated multimodal learner profiles inform future AI–IoT-enabled personalized English-learning systems?

The first six questions can be investigated using the analyses described in the supplied report, subject to verification of the complete statistical outputs. RQ7 represents a potential application that requires additional evaluation.

2.7 Proposed System Architecture

2.71 Architecture overview

The proposed architecture transforms the AI–IoT English Learning Behavior Dataset into interpretable, multidimensional learner profiles. It combines data validation, preprocessing, statistical analysis, dimensionality reduction, clustering, and contextual association modelling.

The architecture is a proposed research framework rather than a claim that a real-time adaptive learning platform has already been implemented.

2.72 Description of the architecture

Data acquisition layer. The study uses the AI–IoT English Learning Behavior Dataset, described in the supplied



Figure 1. Proposed architecture for multimodal learning analytics and English-language proficiency profiling

analysis as containing 12,500 learning-session records from 2,479 unique students. The six primary outcomes are Vocabulary, Grammar, Reading, Writing, Listening, and Speaking. Additional contextual variables include learning activities, attention, engagement, motivation, stress, behavior, learning mode, and device type.

Data-quality and preprocessing layer. The data are checked for missing values, duplicates, identifier integrity, inconsistent categories, and implausible score values. Student-level averages are calculated for the six skill domains to support the principal comparisons. Standardization is applied where required for PCA and clustering.
Analytical layer. This layer contains three complementary analytical pathways:

• Statistical analysis to characterize proficiency distributions, test differences between skills, and investigate relationships among language domains.

• Unsupervised analysis to identify shared proficiency dimensions and exploratory student clusters.

• Regression analysis to investigate adjusted associations between contextual variables and each English-language skill.

Validation layer. Statistical assumptions, effect sizes, confidence intervals, clustering stability, and model diagnostics are examined before interpreting the outputs. Where predictive models are developed, validation must be performed on students excluded from model training.

Learner-profiling layer. The final outputs summarize overall proficiency, domain-specific performance, and within-student imbalances. These profiles may inform instructor review and future personalized support, but should not independently determine a learner's proficiency classification.

3. Methodology

3.1 Research design

This study adopts a quantitative, observational research design within the broader field of educational data mining and multimodal learning analytics. The methodology combines descriptive statistics, repeated measures comparisons, correlation analysis, multivariate testing, dimensionality reduction, clustering, and regression modelling.

The research focuses on two complementary levels of analysis: cohort-level patterns in English language proficiency and student level variation across the six skill domains. This distinction is important because similar average scores do not necessarily imply that individual students have similar learning needs.

The study is observational. Therefore, the analyses identify statistical relationships and exploratory profiles rather than causal effects of learning technologies or instructional interventions.

3.2 Dataset and variables

The supplied report describes 12,500 learning-session records from 2,479 unique students. Its data-quality assessment reports 35 variables, 12,500 unique session identifiers, no missing cells, no exact duplicate rows, and no duplicated session IDs. The dataset description elsewhere states that it contains 34 features; this inconsistency must be reconciled against the original data file and data dictionary.

The primary dependent variables are the six English-language skill scores:

• Vocabulary

• Grammar

• Reading

• Writing

• Listening

• Speaking

The contextual variables described in the manuscript include age, learning activities, IoT-related attention and engagement, gender, grade level, learning mode, device type, stress, motivation, and behavior.

The precise definitions, units, coding schemes, and measurement sources of these variables should be documented before the manuscript is submitted.

3.3 Data preprocessing

The preprocessing stage should include the following procedures.

1. Verify the dataset dimensions, variable types, student identifiers, and session identifiers.

2. Assess missingness, duplicate records, invalid values, inconsistent categories, and score ranges.

3. Aggregate the six language-skill scores across each student's available sessions for the primary student-level comparisons.

4. Retain the number of sessions contributed by each student to assess the potential effect of unequal session counts.

5. Standardize the six skill scores before PCA and K-means clustering.

6. Encode categorical predictors appropriately for regression.

7. Exclude identifiers and variables that directly leak information about the target outcome.

The reported absence of missing cells means that imputation was not required for the screened data. Nevertheless, range checks and measurement validation remain necessary.

Student level aggregation reduces the risk of treating repeated observations from the same learner as independent. However, it also removes information about session to session variation. Where appropriate, session level analyses should be conducted using methods that explicitly account for repeated observations.

3.4 Descriptive and repeated-measures analysis

Descriptive statistics are calculated for each English-language skill, including the mean, standard deviation, median, and distribution.

The reported student level mean scores are:

Skill	Mean score	Standard deviation
Vocabulary	67.82	8.29
Grammar	67.75	8.62
Reading	67.91	8.54
Writing	67.98	8.56
Listening	67.86	8.53
Speaking	67.88	8.77

Repeated measures ANOVA is used to examine whether the mean scores differ across the six skill domains. The Friedman test provides a non parametric sensitivity analysis, while Holm adjusted paired comparisons may be used to investigate individual skill contrasts.

The supplied report gives a repeated measures ANOVA result of $F(5,12390)=0.640$, $p=0.6695$, with

partial $\eta^2=0.0003$. The Friedman test reports $\chi^2(5)=2.104$, $p=0.8346$, and Kendall's $W=0.0002$.

These results do not indicate a statistically significant overall difference among the six skills. However, the reported ANOVA degrees of freedom should be verified against the stated student-level sample size of 2,479 before publication.

3.5 Correlation analysis

Pearson correlation analysis is used to examine the relationships among the six skill scores.

The reported correlations are all positive, ranging from 0.650 to 0.709. The strongest association is between Vocabulary and Reading ($r=0.709$), whereas the weakest is between Writing and Speaking ($r=0.650$).

A correlation heatmap is used to display these relationships. The results suggest that students with relatively high scores in one domain tend to have relatively high scores in other domains. However, the correlations do not establish causation or demonstrate that the skills are interchangeable.

For complete statistical reporting, confidence intervals and appropriate significance tests for the correlations should also be supplied.

3.6 Within-student proficiency gaps

To examine individual differences, the highest and lowest skill scores are identified for each student.

The within-student gap is defined as:

$$ G_i=\max(S_{i1},\ldots,S_{i6})- \min(S_{i1},\ldots,S_{i6}) $$

where S_{ij} represents the mean score of student i in skill domain j.

The reported mean gap is 11.85 points, while the median is 10.94 points. These statistics indicate that individual skill differences can be more substantial than the small differences between cohort wide means.

The strongest and weakest skill labels are interpreted as relative rankings within each student. They do not constitute absolute proficiency classifications because the report does not establish an external proficiency benchmark.

3.7 Multivariate group comparisons

Multivariate analysis of variance (MANOVA) is used to examine whether the combined six-skill profile differs across gender, grade level, and learning mode.

Wilks' lambda and Pillai's trace are reported as the principal omnibus statistics. Appropriate univariate tests and adjusted post-hoc comparisons should follow when justified by the omnibus results.

The supplied results do not show consistent statistically significant differences in the combined skill profiles by gender, grade level, or learning mode under Wilks' lambda and Pillai's trace. An isolated significant result under Roy's greatest root statistic for grade level should be interpreted cautiously.

The report also describes Welch's independent samples t-test, one-way ANOVA, and Kruskal Wallis tests for relevant group comparisons. Exact group specific effects should not be asserted until the corresponding output tables are available.

3.8 Principal component analysis

PCA is performed on standardized skill scores to investigate whether the six domains share a dominant

dimension of variation.

The reported first principal component has an eigenvalue of 4.3866 and explains 73.08% of standardized variance. The first two components together explain 79.19%.

These findings indicate that a large proportion of the variation in the six observed scores is shared. Nevertheless, component loadings are required to determine how strongly each skill contributes to the components.

The reported KMO statistic of 0.931 and Bartlett's test result of $\chi^2(15)=10370.80$, $p<0.001$, support examining the correlation structure further. They do not, by themselves, establish the validity of a psychological or educational measurement model.

3.9 Student-profile clustering

K-means clustering is used to explore similarities among students based on their standardized six-skill profiles. Candidate solutions with k values from 2 to 6 are compared using silhouette scores.

The supplied analysis selects a two-cluster solution with a silhouette score of 0.353. The reported mean profiles are nearly parallel, with one cluster scoring higher across all six skills.

This pattern suggests that the solution primarily separates students by overall score level rather than by distinct combinations of skill-specific strengths. Consequently, the clusters should be considered exploratory. Further assessment should include cluster sizes, repeated initializations, resampling-based stability checks, alternative algorithms, and sensitivity to feature scaling.

3.10 Skill-specific regression analysis

Separate ordinary least squares models are described for Vocabulary, Grammar, Reading, Writing, Listening, and Speaking.

The candidate predictors include age, learning activities, IoT-related attention and engagement, gender, grade level, learning mode, device type, stress, motivation, and behavior.

The general model for skill j is:

$$ S_{ij}=\beta_{oj}+\sum_{k=1}^{p}\beta_{kj}X_{ik} +\varepsilon_{ij} $$

where:

- S_{ij} is the score of student i in skill j.
- X_{ik} represents predictor k.
- β_{oj} is the intercept for skill j.
- β_{kj} is the adjusted coefficient for predictor k.
- ε_{ij} is the residual term.

The source describes HC3 heteroskedasticity-robust standard errors. Quiz scores, assignment scores, and overall learning-performance scores are excluded to reduce potential circularity between assessment-based predictors and the skill outcomes.

The final manuscript should report coefficients, confidence intervals, p-values, adjusted R^2, sample sizes, residual diagnostics, and multicollinearity checks. Since these complete outputs are absent from the supplied report, the direction and magnitude of individual predictor associations cannot yet be established.

4. Experimental Setup

4.1 Analytical environment

The experimental workflow should be implemented in a reproducible statistical environment capable of supporting data cleaning, descriptive statistics, repeated-measures testing, PCA, clustering, regression, and visualization.

The supplied manuscript does not specify the software versions, package versions, or operating environment. These details should be reported from the actual analysis environment rather than inferred.

4.2 Experimental workflow

The analysis should proceed in the following order:

1. Dataset inspection and data-quality assessment.

2. Student-level aggregation of the six language skills.

3. Descriptive statistics and proficiency-distribution visualizations.

4. Repeated-measures ANOVA and Friedman testing.

5. Correlation analysis and heatmap construction.

6. Within-student gap analysis and relative strength/weakness identification.

7. MANOVA and relevant subgroup comparisons.

8. PCA and factorability assessment.

9. K-means clustering and cluster-stability assessment.

10. Skill-specific OLS regression with HC3 robust standard errors.

11. Statistical diagnostics, sensitivity analyses, and reconciliation of results.

4.3 Validation strategy

The existing analysis is predominantly descriptive and inferential; it is not a completed predictive-model benchmark.

If future experiments develop models to predict continuous learning performance scores or categorical learner behaviors, the data should be partitioned by student ID, rather than randomly by session. Otherwise, records from the same student could appear in both training and test data, leading to overly optimistic evaluation.

A proposed 70% training, 15% validation, and 15% test split could be used for such future predictive experiments. These proportions are recommendations, not reported experimental settings.

All learned preprocessing operations, feature selection, and hyperparameter tuning should be confined to the training and validation procedures. Final performance should be reported on a held out test set.

For continuous predictions, suitable metrics include MAE and RMSE. For categorical predictions, appropriate metrics include macro-F1, balanced accuracy, and confusion matrices. These metrics should only be reported after the corresponding models have been trained and evaluated.

5. Analysis

The dataset contains 12,500 learning-session records from 2,479 unique students and 35 variables. The six English-language dimensions are Vocabulary, Grammar, Reading, Writing, Listening, and Speaking. Because students contribute repeated sessions, the principal comparisons of language skills use the student-level mean across each student's available sessions. This makes the student, rather than each session, the unit of analysis for the main skill profile comparisons. It reduces the risk of treating multiple observations from the same learner as independent, although averaging also removes information about changes between that learner's sessions.

5.1 Data quality assessment

Metric	Value
Session records	12,500
Variables	35
Unique students	2,479
Unique sessions	12,500
Missing cells	0
Exact duplicate rows	0
Duplicate session IDs	0
Students with repeated sessions	10,021

Table 1. Data quality and dataset structure

Table 1 summarizes the size, completeness, and identifier integrity of the dataset. It contains 12,500 session records and 35 variables, contributed by 2,479 distinct students. All 12,500 session identifiers are unique. The data quality screening identified no missing cells, no exact duplicate rows, and no duplicated session identifiers. A total of 10,021 students have repeated session records.

The absence of missing values and exact duplicate records provides a clean starting point for the planned analyses and means that missing value imputation or exact row deduplication is not required for the variables screened. The repeated session structure is nevertheless important: 10,021 students contribute more than one session, so session level observations are not statistically independent. Analyses that compare the six language skills should therefore account for student level dependence, as the main report does by averaging scores within each student. The table does not, by itself, establish that all measurements are valid, correctly scaled, or free from unusual values; range checks, outlier checks, and verification of score generation procedures remain relevant.

5.2 Descriptive English proficiency profile

Table 2 reports the student level mean, standard deviation, median, and rank for each English-language skill. Writing has the highest mean (67.98), followed closely by Reading (67.91), Speaking (67.88), Listening (67.86), Vocabulary (67.82), and Grammar (67.75). The standard deviations range from 8.29 for Vocabulary to 8.77 for Speaking, indicating broadly comparable dispersion across the six domains. The medians are also close to the corresponding means.

Skill	Mean	SD	Median	Rank
Vocabulary	67.82	8.29	67.66	5
Grammar	67.75	8.62	67.90	6
Reading	67.91	8.54	67.85	2
Writing	67.98	8.56	67.98	1
Listening	67.86	8.53	67.88	4
Speaking	67.88	8.77	68.22	3

Table 2. Descriptive statistics and mean-based ranking of the six English skills

The descriptive ranking suggests that the cohort's average performance is relatively balanced across the six language domains. Writing is numerically the highest-scoring skill and Grammar the lowest, but the difference between these two means is only 0.23 points. This small descriptive difference should not be interpreted as evidence that Writing is meaningfully stronger or Grammar meaningfully weaker at the population level; the paired inferential tests reported later do not detect a statistically significant overall difference among the six domains. The similar standard deviations suggest that variability between students is of a comparable order across skills. The reported mean individual maximum to minimum gap of 11.85 points (median 10.94) is a different quantity: it describes within student imbalance and can be appreciable even when cohort wide skill means are nearly identical.

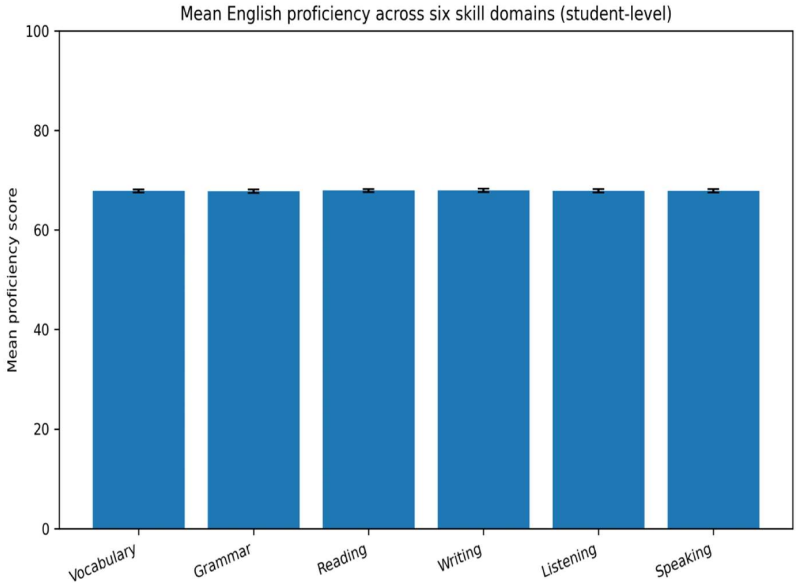

Figure 2. Mean student-level proficiency scores with approximate 95% confidence intervals

The bar chart displays the average student-level score for each of the six English skills. Vertical error bars represent approximate 95% confidence intervals around the means. The six bars are tightly grouped near 68 points, and the confidence intervals are narrow because the estimates are based on 2,479 student profiles.

The figure visually confirms the close similarity of average performance across the six skill domains. Writing is marginally highest and Grammar marginally lowest, but the differences are small relative to the overall score scale. The narrow intervals indicate that the cohort means are estimated with reasonable precision; they

do not imply that individual students have little variability. This is why the distribution plot and within-student gap analysis are also important. The overlapping intervals are consistent with, but do not replace, the formal repeated measures tests.

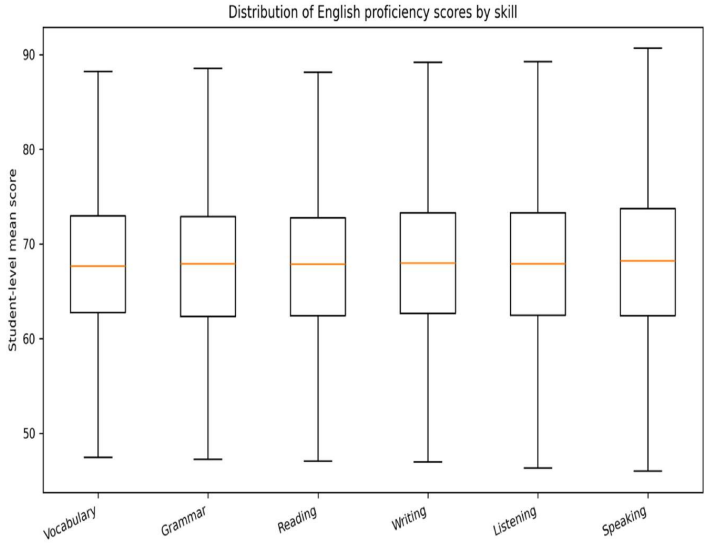

Figure 3. Distribution of student-level mean scores by English skill

The boxplots compare the distributions of student-level mean scores for Vocabulary, Grammar, Reading, Writing, Listening, and Speaking. Each box represents the interquartile range, the line inside the box is the median, and the whiskers show the displayed range after excluding extreme points for visual clarity.

The medians and interquartile ranges are similar across the six domains, supporting the conclusion that the overall proficiency distributions are broadly comparable. The boxes also show meaningful variation between learners within each skill. Therefore, similar cohort averages should not be taken to mean that all students have the same proficiency pattern. The figure is descriptive and does not test whether paired skill scores differ statistically.

5.3 Paired comparisons among the six skills

The repeated-measures ANOVA reports $F(5, 12{,}390) = 0.640$, $p = 0.6695$, with partial eta-squared = 0.0003. The Friedman test reports $\div^2(5) = 2.104$, $p = 0.8346$, and Kendall's W = 0.0002. Both tests are based on the same students' six skill scores, and both assess whether the central tendency differs across the within-student skill conditions.

Neither the parametric repeated-measures test nor the non-parametric Friedman test provides evidence of an overall difference among Vocabulary, Grammar, Reading, Writing, Listening, and Speaking. The extremely small partial eta-squared and Kendall's W indicate that the observed skill domain effect is negligible in this dataset. Consequently, the numerical ranking in Table 2 should be described as a descriptive ordering rather than a statistically established hierarchy. The Holm adjusted pairwise Wilcoxon results are referenced in the source document as Table 9, but that table is not included in the supplied 10-page file; individual pairwise claims therefore cannot be made from this version of the report.

Each faint line represents the six skill scores for one student in a displayed sample. The bold line shows the cohort mean for each skill. The student traces vary substantially around the overall mean, while the cohort-average line remains almost flat.

The figure illustrates why similar group means can coexist with meaningful individual differences. Some students show relative strengths in particular skills and weaknesses in others, whereas other students have more even

profiles. The near-horizontal overall mean line is consistent with the non-significant repeated-measures tests. Because only a sample of student traces is displayed, the figure should be used to illustrate profile heterogeneity rather than to estimate the prevalence of any particular profile.

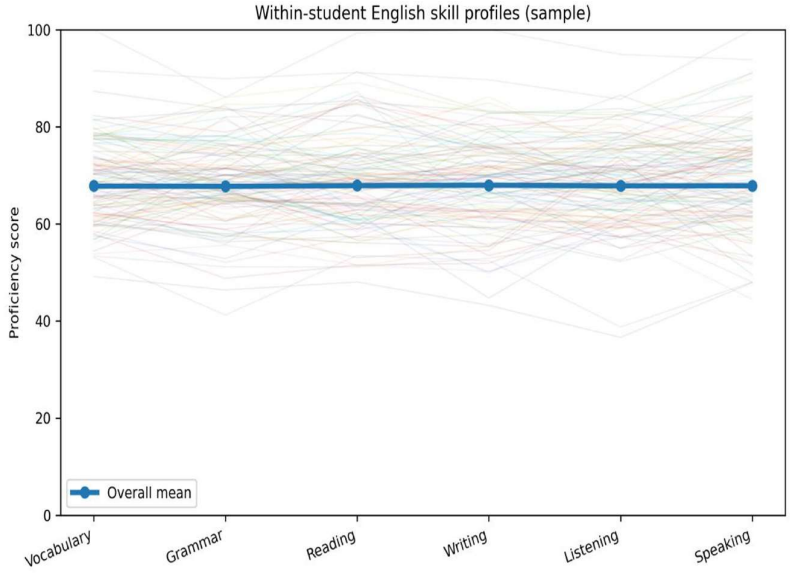

Figure 4. Within-student skill profiles for a sample of students, overlaid with the overall mean

5.4 Correlation among English language skills

Skill	Vocabulary	Grammar	Reading	Writing	Listening	Speaking
Vocabulary	1.000	0.699	0.709	0.664	0.696	0.679
Grammar	0.699	1.000	0.691	0.651	0.678	0.671
Reading	0.709	0.691	1.000	0.652	0.699	0.676
Writing	0.664	0.651	0.652	1.000	0.666	0.650
Listening	0.696	0.678	0.699	0.666	1.000	0.672
Speaking	0.679	0.671	0.676	0.650	0.672	1.000

Table 3. Pearson correlation matrix for the six English language skill

Table 3 presents pairwise Pearson correlation coefficients among the six student level skill scores. All off-diagonal coefficients are positive and range from approximately 0.650 to 0.709. The strongest association is between Vocabulary and Reading (r = 0.709), followed by Reading and Listening (r = 0.699) and Vocabulary and Grammar (r = 0.699). The weakest displayed association is between Writing and Speaking (r = 0.650). The diagonal values equal 1 because each skill is perfectly correlated with itself.

The uniformly positive correlations indicate that students who perform relatively well in one English domain also tend to perform relatively well in the other domains. This pattern is consistent with a shared general language proficiency component, while the coefficients remain below 1.0, leaving room for skill specific

strengths and weaknesses. The Vocabulary Reading association is plausible as a statistical relationship because lexical knowledge and reading performance may covary, but the correlation alone does not establish a learning mechanism or causal direction. The matrix contains coefficients but not their p-values or confidence intervals; significance and uncertainty should be taken from the separate pairwise correlation test output if included in the full analysis package.

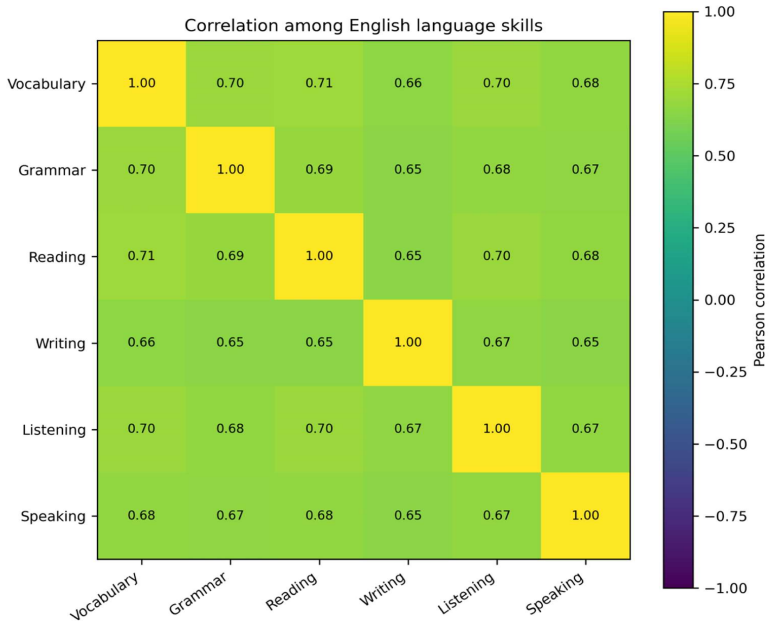

Figure 5. Pearson correlation heatmap across six skill domains

The heatmap visualizes the Pearson correlation matrix. The diagonal is 1.00, and the off-diagonal cells are labelled with their corresponding correlation coefficients. Similar positive values appear throughout the matrix, with the strongest cell corresponding to Vocabulary and Reading.

The consistent positive colouring demonstrates that the six skill scores are related rather than independent. The heatmap supports investigating an overall proficiency dimension, which is examined further through PCA. At the same time, the correlations are not so close to 1.00 that the six domains are interchangeable. Each skill should therefore remain visible in descriptive reporting and in student level diagnostic profiles.

5.5 Relative strengths and within-student proficiency gaps

For each student, the highest of the six mean skill scores was labelled the strongest skill and the lowest was labelled the weakest skill. These are relative within student labels, not externally norm-referenced proficiency categories. The mean individual maximum to minimum gap is 11.85 points and the median gap is 10.94 points.

The grouped bars show, for each skill, the percentage of students for whom that skill has the highest score and the percentage for whom it has the lowest score. The strongest-skill percentages are fairly evenly distributed, with Writing appearing most frequently as the strongest domain. Grammar appears most frequently as the weakest domain, while the other skills also occupy both positions for substantial shares of students.

The distribution indicates that no single English skill is universally strongest or weakest across the student population. Although Writing has the highest cohort mean and Grammar the lowest, individual students exhibit different patterns. This supports differentiated learning support rather than assuming that every learner needs the same intervention. The labels are relative: a student's strongest skill could still be low in absolute terms, and the weakest skill could still be adequate. The chart should therefore not be used to classify students as proficient or deficient without an external benchmark or validated proficiency threshold.

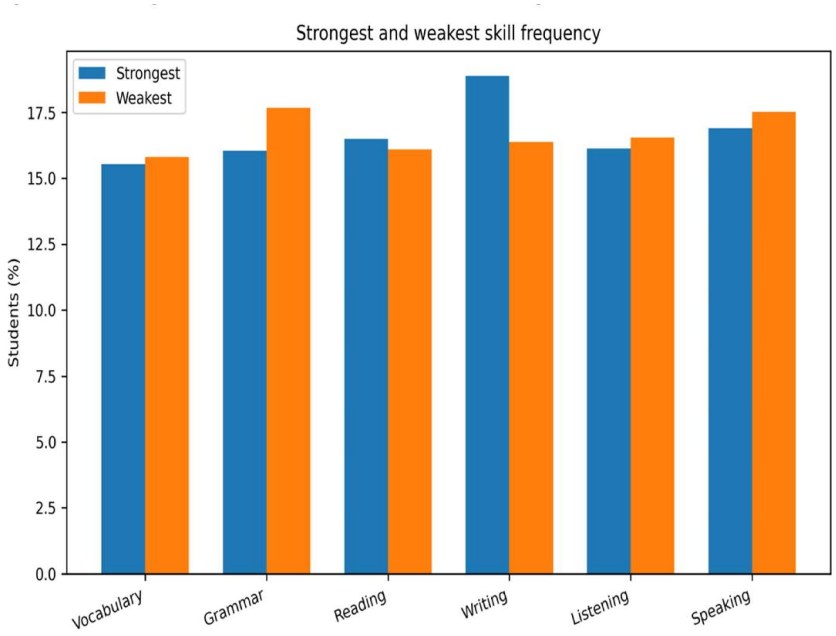

Figure 5. Percentage of students for whom each domain is strongest or weakest

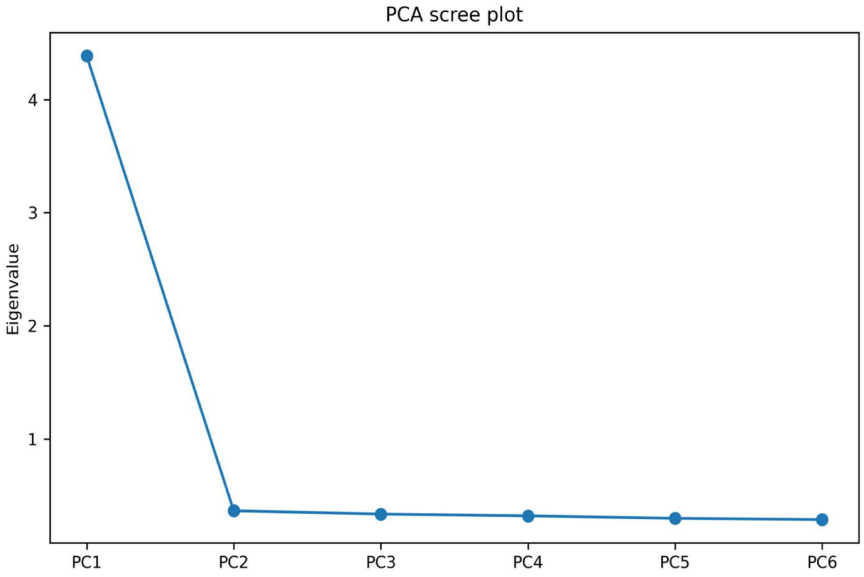

Figure 6. Distribution of within-student maximum to minimum skill gaps

The histogram displays the difference between each student's highest and lowest mean English skill score. The horizontal axis represents the size of this gap in score points, and the vertical axis shows the number of students in each interval. Most students are concentrated in the lower to middle gap range, while a smaller number have much larger differences.

The distribution shows that a typical student has a noticeable but not extreme difference between their best and least developed skill, with the median gap at 10.94 points. The right tail indicates that some learners have substantially more uneven profiles and may benefit from targeted skill level assessment. These larger gaps are potential flags for follow up, not proof of a learning difficulty. The analysis should be supplemented by examining

which skill is responsible for each student's gap and whether the gap is stable across repeated sessions.

5.51 Group differences in English proficiency

The source analysis describes skill specific comparisons by Gender, Grade_Level, Learning_Mode, and Device_Type. Welch's independent samples t-test was used for two-level comparisons, and one-way ANOVA for comparisons with more than two groups. Kruskal–Wallis tests were used as non parametric sensitivity analyses; Holm-adjusted p-values, effect sizes, and Tukey post-hoc comparisons were also generated. The detailed tables for these tests (identified in the source as Tables 13–14) are not present in the supplied document, so exact group-level test statistics and pairwise differences cannot be described here.

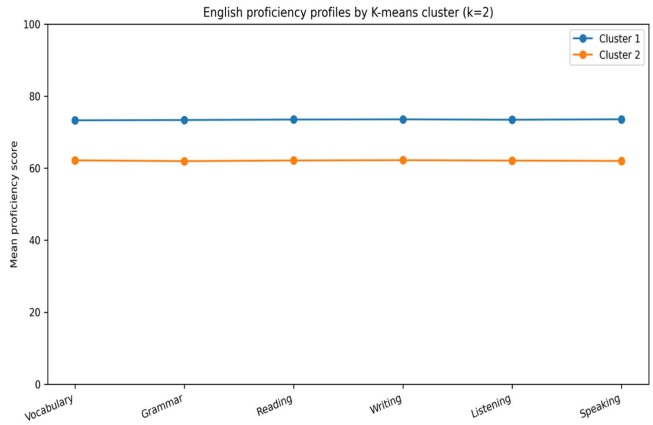

Figure 7. Mean skill scores across grade levels

The line chart plots the mean score for each of the six English skills across the grade categories shown on the horizontal axis. The six lines generally follow a similar pattern: mean scores vary modestly across the earlier grade categories, rise noticeably at Grade 8, and then decline to varying degrees at Grade 9. The vertical scale is restricted to a narrow band around the observed means, so small absolute differences are visually emphasized.

The chart suggests that grade-level differences may not be uniform across the six skill domains, although the broad shape is similar for most skills. Grade 8 has the highest plotted mean for all or nearly all domains, while Grade 9 shows lower means than Grade 8. These are descriptive patterns and should not be interpreted as a developmental trajectory because the dataset does not provide verified longitudinal tracking of the same students across grades. The multivariate results in Table 4 also do not provide consistent evidence of a grade-level effect across the commonly used omnibus statistics. Any grade-related claim should be based on the full univariate and post-hoc results, including effect sizes and adjusted p-values.

7. Multivariate skill profile differences (MANOVA)

Effect	Statistic	Value	F	Numerator df	Denominator df	p
Gender	Wilks' lambda	0.99898	0.4193	6	2465	0.8666
Grade level	Wilks' lambda	0.98727	1.3190	24	8600.568	0.1362
Grade level	Pillai's trace	0.01277	1.3170	24	9872	0.1374
Learning mode	Wilks' lambda	0.99312	0.9473	18	6972.558	0.5197
Learning mode	Pillai's trace	0.00690	0.9477	18	7401	0.5191

Table 4. Selected MANOVA statistics for multivariate English skill profiles

Table 4 reports multivariate tests of whether the combined six-skill profile differs according to Gender, Grade_Level, and Learning_Mode. Several equivalent MANOVA statistics are shown in the source output: Wilks' lambda, Pillai's trace, Hotelling–Lawley trace, and Roy's greatest root. The intercept rows test whether the multivariate response mean differs from zero and are not the substantive group comparisons of interest. The gender effect is non-significant (Wilks' lambda = 0.99898, F = 0.419, p = 0.8666). The grade-level effect is also non-significant under Wilks' lambda (p = 0.1362) and Pillai's trace (p = 0.1374). The learning-mode effect is non-significant under Wilks' lambda (p = 0.5197) and Pillai's trace (p = 0.5191).

Using the conventional omnibus criteria of Wilks' lambda and Pillai's trace, the reported MANOVA does not show statistically significant differences in the combined English skill profile by gender, grade level, or learning mode. The Roy's greatest root statistic for grade level is reported as significant in the original output (p = 0.000446), whereas the other three grade-level statistics are not. Because the MANOVA criteria can differ in sensitivity and Roy's root is especially sensitive to a dominant single dimension, this isolated result should be treated cautiously rather than presented as unambiguous evidence of a multivariate grade effect. The small values of the reported Pillai's trace and the non-significant p-value are consistent with a limited overall group separation. Device type is not included in the displayed MANOVA table, even though it is mentioned in the broader group-comparison analysis. Follow up univariate tests and adjusted post-hoc comparisons are needed before identifying any particular skill or grade contrast.

5.6 Principal component analysis and factorability

Component	Eigenvalue	Explained variance	Cumulative variance
PC1	4.3866	73.08%	73.08%
PC2	0.3669	6.11%	79.19%
PC3	0.3377	5.63%	84.82%
PC4	0.3222	5.37%	90.19%
PC5	0.3000	5.00%	95.18%
PC6	0.2891	4.82%	100.00%

Table 5. PCA eigenvalues and explained variance for the six English skill scores

Table 5 summarizes the variance accounted for by each principal component extracted from the six standardized skill scores. The first component (PC1) has an eigenvalue of 4.3866 and explains 73.08% of the total standardized variance. PC2 explains a further 6.11%, bringing the cumulative variance explained by the first two components to 79.19%. The remaining four components each explain between approximately 4.82% and 5.63%.

The dominance of PC1 indicates that a large proportion of the variation in the six skill scores is shared across domains. This is consistent with the positive correlations in Table 3 and suggests that an overall English proficiency dimension may summarize much of the common variation. The first two components together retain 79.19% of the standardized variance, providing a compact two-dimensional representation for exploratory visualization. However, the component variance percentages alone do not show which skills contribute most strongly to each component; component loadings are needed for that interpretation. PC1's large eigenvalue and the sharp drop to PC2 suggest a dominant general dimension, but PCA is a data-reduction

method and does not, by itself, validate a psychological or educational measurement model.

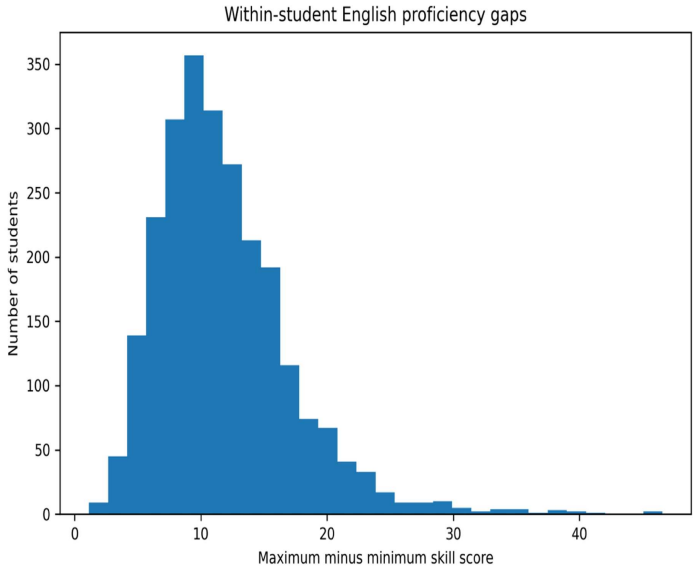

Figure 8. PCA projection of student proficiency profiles

Each point represents one student projected onto the first two principal components calculated from standardized six-skill scores. Most observations form a dense central cloud, with a smaller number of students located farther from the centre.

The dense central concentration indicates that many students have broadly similar overall proficiency profiles, while the peripheral points indicate learners whose combined skill patterns differ more strongly from the cohort centre. The plot is useful for visualizing multivariate variation and potentially unusual profiles, but the axes are combinations of the original skills rather than individual skill scores. Without loading vectors or labelled groups, the figure cannot identify which specific skill explains a student's position or establish naturally separated subpopulations.

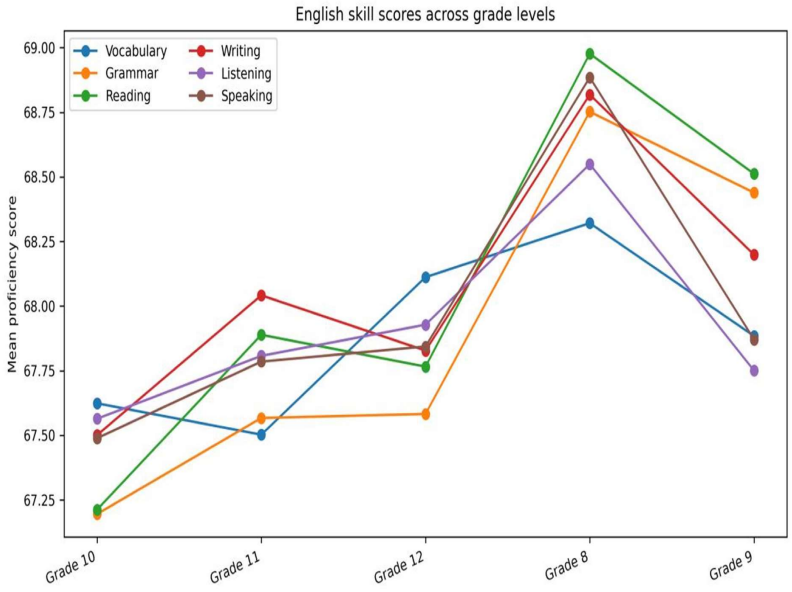

Figure9. PCA scree plot

The scree plot displays the eigenvalue associated with each of the six principal components. There is a pronounced drop from PC1 to PC2, followed by a much flatter sequence for PC2 through PC6.

The sharp elbow after PC1 indicates that the first component captures substantially more variation than any subsequent component. This supports the interpretation of a dominant shared proficiency dimension. The additional components each contribute relatively small portions of variance. The scree plot is a descriptive aid for deciding how many components to retain; the final decision should also consider the research purpose, component loadings, interpretability, and validation in an independent sample.

Factorability diagnostics. The source report gives an overall KMO value of 0.931 and Bartlett's test ÷²(15) = 10,370.80, with the reported p-value displayed as 0 (which should be written as $p < 0.001$, not as a literal zero). These results indicate that the correlation matrix is sufficiently interrelated for exploratory factor analysis to be considered. Nevertheless, only six indicators are available, and the two factor solution is described in the source as a diagnostic. Factor loadings, extraction details, rotation, and factor retention evidence should be reviewed before making claims about distinct latent language constructs.

5.8 Student proficiency-profile clustering

K-means solutions with k values from 2 to 6 were compared using silhouette scores. The selected solution was k = 2, with a silhouette score of 0.353. The figure shows two nearly parallel skill profiles: Cluster 1 has a higher mean across all six skills than Cluster 2.

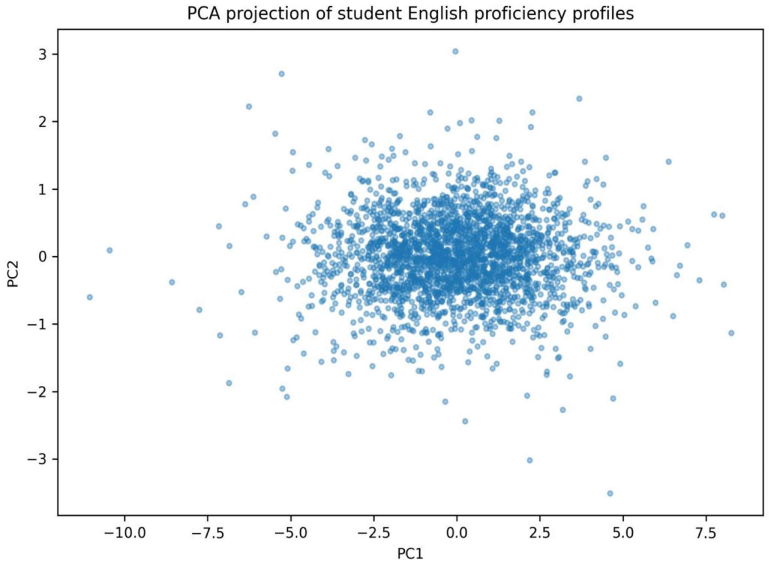

Figure 10. Mean English skill profile for each selected cluster

The line chart compares the mean Vocabulary, Grammar, Reading, Writing, Listening, and Speaking scores for the two selected K-means clusters. Cluster 1 is consistently above Cluster 2 across all six domains. The lines are nearly parallel, suggesting that the main distinction between clusters is overall score level rather than a sharply different pattern of relative strengths across skills.

The selected clustering appears to separate students into relatively higher- and lower-scoring overall proficiency groups. Because the profiles are parallel, the solution does not appear to identify distinct skill-specialization types (for example, a group strong in speaking but weak in writing). A silhouette score of 0.353 indicates some cluster structure but not exceptionally strong separation; the clusters should therefore be treated as exploratory summaries rather than definitive student categories. The cluster sizes, full numerical profiles, and stability checks are required before using the groups for educational decisions. Cluster membership is not a diagnosis and should not be used as a substitute for direct assessment.

5.9 Adjusted associations with learning context

The source document states that separate ordinary least squares models were fitted for each of the six English skills. Predictors included age, learning activities, IoT attention and engagement, gender, grade level, learning mode, device type, stress, motivation, and behaviour. HC3 heteroskedasticity robust standard errors were used. Quiz scores, assignment scores, and overall learning-performance scores were excluded to reduce circularity between the language outcomes and other assessment-based measures.

This modelling strategy is intended to estimate the adjusted association between each learning-context variable and a particular English skill while accounting for the other included predictors. Fitting a separate model for each skill allows the associations to differ across Vocabulary, Grammar, Reading, Writing, Listening, and Speaking rather than assuming that all domains respond identically to the same learning conditions. HC3 robust standard errors provide some protection against heteroskedasticity in the residuals. However, the supplied 10-page document does not display the regression coefficient table, confidence intervals, p-values, model R-squared values, or sample sizes. Therefore, it is not possible to identify which predictors are statistically associated with each skill, the direction or magnitude of those associations, or how much variance the models explain. Those claims should be added only after the actual regression output is inserted. Even when coefficients are statistically significant, they remain adjusted observational associations and should not be interpreted as causal effects.

5.10 Expanded limitations and implications for interpretation

Observational design and causal interpretation. The dataset is observational and records student characteristics, learning activities, IoT-related indicators, and proficiency scores without a randomized intervention. Consequently, associations between attention, engagement, motivation, learning mode, or activity patterns and English scores cannot establish that these factors caused changes in proficiency. Reverse relationships and unmeasured confounding are possible; for example, students with stronger prior proficiency may engage differently with learning activities. Causal claims would require an appropriate experimental or quasi-experimental design, temporal ordering, and stronger control of confounding.

Absence of verified timestamps and session sequence. The supplied data do not contain a verified timestamp or reliable chronological session sequence. It is therefore not possible to determine whether a student's vocabulary, grammar, reading, writing, listening, or speaking score improved or declined over time. Differences between students or between their average session scores must not be described as learning gains, growth rates, or longitudinal progress. A future dataset should retain timestamped sessions and consistent repeated assessments so that within-student change can be modelled.

Aggregation of repeated sessions. The main skill comparisons use the mean score across each student's available sessions. This student level aggregation prevents students with many sessions from being counted as many independent learners in the primary comparisons. However, it also removes session to session variation and may conceal short term changes, fluctuations, or context-specific performance. Students may contribute different numbers of sessions, and their averages may therefore differ in precision. Future work could use multilevel mixed effects models or generalized estimating equations with the session level data, provided that session ordering and relevant covariates are available.

Interpretation of relative strongest and weakest skills. The strongest and weakest skill labels are calculated by comparing the six scores within each student. They indicate relative ranking only. A student's highest score may still fall below an appropriate proficiency standard, while the lowest score may still represent adequate performance. The dataset does not provide an external benchmark such as CEFR levels or a validated criterion score. Future reporting should distinguish relative within-student strengths from absolute proficiency classification and should use validated cut offs if proficiency bands are required.

Exploratory nature of PCA and EFA. PCA and EFA are exploratory analyses of the relationships among six observed skill scores. The high KMO value and significant Bartlett test support factorability, but they do not establish construct validity. PCA components are mathematical combinations that maximize explained

variance, whereas latent factors require a defensible measurement interpretation. With only six indicators, factor retention, loading stability, and construct differentiation require caution. A stronger validation study should examine factor loadings, use an appropriate rotation where justified, test measurement reliability, and confirm the proposed structure in an independent sample using confirmatory factor analysis.

Exploratory nature and stability of clustering. The two-cluster K-means solution has a silhouette score of 0.353, indicating moderate or limited separation rather than sharply distinct groups. K-means results depend on scaling, initialization, the selected number of clusters, and the variables included. The nearly parallel cluster profiles suggest that the main separation is overall proficiency level. Before translating these groups into personalized interventions, the solution should be tested using repeated initialization, bootstrap or subsampling stability, alternative clustering algorithms, and sensitivity to scaling and cluster number. The clusters should remain descriptive profiles, not fixed learner identities.

Statistical significance, effect sizes, and sample size. The student sample is relatively large, so small effects may be estimated precisely and may become statistically significant even when their practical importance is limited. Conversely, a non-significant test does not prove that two groups are exactly equivalent. Interpretation should consider effect sizes, confidence intervals, score scale meaning, and educational relevance in addition to p-values. Where multiple skills or group contrasts are tested, multiplicity adjustments such as Holm correction should be retained and clearly reported.

Measurement validity and score comparability. The analysis treats the six skill scores as comparable numerical measures. This interpretation assumes that the scores use compatible scales and that scoring procedures are sufficiently consistent across students, sessions, and skill domains. The supplied document does not establish the assessment instruments' reliability, rubric validity, inter rater agreement, or measurement invariance across gender, grade, or learning mode. These properties should be documented before strong conclusions are drawn about small differences between domains or groups.

Incomplete supporting outputs in the supplied report. The source document refers to additional pairwise Wilcoxon results, group-comparison and post-hoc tables, clustering profile tables, and adjusted regression coefficients that are not reproduced in the supplied 10-page file. As a result, the present expanded interpretation can describe the reported summary statistics and displayed figures but cannot responsibly state the exact pairwise differences, group specific effects, cluster sizes, or regression predictor effects. These tables should be inserted from the computed analysis outputs before the report is submitted as a complete empirical results section.

6. Overall synthesis

Taken together, the displayed results indicate that the cohort has closely similar average scores across the six English skill domains, with no statistically significant overall difference detected by either repeated measures ANOVA or the Friedman test. At the same time, the individual-level distribution and proficiency gap analyses show that students differ in their relative skill profiles. All six domains are positively correlated, and PCA identifies a dominant common dimension that accounts for most of the standardized variance. The exploratory two cluster solution appears to separate higher- and lower-scoring profiles more than it distinguishes domain-specific learning styles. The MANOVA results do not show consistent omnibus evidence of skill-profile differences by gender, grade, or learning mode under Wilks' lambda and Pillai's trace. Finally, the adjusted regression approach is appropriate in principle for exploring contextual associations, but its coefficients must be included before substantive conclusions can be drawn about the contribution of IoT indicators, learning activities, or behavioural factors. The strongest educational implication supported by the current evidence is therefore the value of assessing individual skill profiles rather than relying only on cohort wide averages.

7. Conclusion

This study demonstrates the viability of leveraging AI–IoT datasets and Multimodal Learning Analytics (MMLA)

to construct multidimensional profiles of English language learners. The primary contribution of this research is the empirical evidence that cohort wide average scores can mask significant individual learning needs. While the six English language skills are highly correlated and largely driven by a shared general proficiency dimension, individual students exhibit distinct intra personal skill imbalances (e.g., relative strengths in Writing versus weaknesses in Grammar).

From an educational perspective, these findings challenge the "one size fits all" approach to language instruction. The exploratory clustering and multivariate analyses reveal that demographic or contextual variables (such as gender, grade level, and learning mode) do not uniformly dictate a student's skill profile. Therefore, adaptive learning systems must move beyond aggregate metrics and utilize continuous, multimodal data to identify and address individual skill specific deficits and strengths.

Despite these insights, this study is bounded by several limitations that define the trajectory for future research. First, the observational nature of the dataset precludes causal inferences regarding how IoT-mediated attention, engagement, or specific learning activities directly impact skill acquisition. Second, the absence of verified chronological timestamps prevents the modeling of longitudinal learning growth or session to session fluctuations. Third, the identification of "strongest" and "weakest" skills is relative to the individual student rather than anchored to absolute, externally validated proficiency benchmarks (e.g., CEFR levels).

Future work should focus on integrating timestamped, longitudinal session data to track learning trajectories over time using multilevel mixed effects models. Additionally, subsequent studies must prioritize the validation of measurement instruments, establish external proficiency criteria, and develop predictive machine learning models trained on held out test sets to translate these exploratory multimodal profiles into real time, actionable personalized interventions in smart educational environments.

References

[1] Wang, Y., Lai, Y., Huang, X. (2024). Innovations in online learning analytics: A review of recent research and emerging trends. *IEEE Access, 12*, 166761–166775. https://doi.org/10.1109/ACCESS.2024.3493621.

[2] Siemens, G., Baker, R. S. J. D. (2012, April). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (p. 252–254).

[3] Guerrero-Sosa, J. D. T., Romero, F. P., Menéndez-Domínguez, V. H., Serrano-Guerrero, J., Montoro-Montarroso, A., Olivas, J. A. (2025). A comprehensive review of multimodal analysis in education. *Applied Sciences, 15*(11), Article 5896. https://doi.org/10.3390/app15115896.

[4] Yang, K., Borchers, C., Falhs, A. C., Echeverria, V., Karumbaiah, S., Rummel, N., Aleven, V. (2024, September 16–20). Leveraging multimodal classroom data for teacher reflection: Teachers' preferences, practices, and privacy considerations. In *Lecture Notes in Computer Science: Vol. 15159. Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics* (pp. 498–511). Krems, Austria.

[5] Lin, H. (2024). Online professional-creative fusion music major students' classroom state recognition based on the integration of DSC and LeNet-5 models. *Journal of Artificial Intelligence and Technology, 4*, 216–226.

[6] Echeverria, V., Yan, L., Zhao, L., Abel, S., Alfredo, R., Dix, S., Jaggard, H., Wotherspoon, R., Osborne, A., Buckingham Shum, S., et al. (2024, March 18–22). TeamSlides: A multimodal teamwork analytics dashboard for teacher-guided reflection in a physical learning space. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 112–122). Kyoto, Japan.

[7] Hakami, L., Hernandez-Leo, D., Amarasinghe, I., Sayis, B. (2024). Investigating teacher orchestration load

in scripted CSCL: A multimodal data analysis perspective. *British Journal of Educational Technology, 55*, 1926–1949.

[8] Ding, A. C. E., Huang, K. T. T., DuBois, J., Fu, H. (2024). Integrating immersive virtual reality technology in scaffolded game-based learning to enhance low motivation students' multimodal science learning. *Educational Technology Research and Development, 72*, 2083–2102.

[9] Na, H., Sung, H. (2025). Learn math through motion: A technology-enhanced embodied approach with augmented reality for geometry learning in K-12 classrooms. *Interactive Learning Environments*. Advance online publication.

[10] Chan, R. Y. Y., Wong, C. M. V., Yum, Y. N. (2023). Predicting behavior change in students with special education needs using multimodal learning analytics. *IEEE Access, 11*, 63238–63251.

[11] Yürüm, O. R. (2025). Technology-enhanced multimodal learning analytics in higher education: A systematic literature review. *IEEE Access, 13*, 92057–92073. https://doi.org/10.1109/ACCESS.2025.3572467

[12] Wang, X. (2025). AI-driven learning behavior analysis and modeling framework for English education based on IoT. *Nonlinear Theory and Its Applications, IEICE, 16*(4), 925–945.

[13] Xu, B. (2025, June). IoT-based multimodal learning framework for predicting student engagement in English education. In *Proceedings of the Second International Conference on Intelligent Transportation and Smart Cities (ICITSC 2025)* (Vol. 13682, p. 872). SPIE.

[14] Chen, Y., Huang, C., Gao, S., Lyu, Y., Chen, X., Liu, S., Bao, D., Lv, C. (2025). A multimodal deep learning approach for legal English learning in intelligent educational systems. *Sensors, 25*(11), Article 3397. https://doi.org/10.3390/s25113397.

[15] Yun, L. (2024, September 22–24). Application of English semantic understanding in multimodal machine learning. In *Proceedings of the 2024 International Conference on Artificial Intelligence, Deep Learning and Neural Networks (AIDLNN)* (p. 233–237). Guangzhou, China.

[16] Bandyopadhyay, D., Hasanuzzaman, M., & Ekbal, A. (2024). Seeing through VisualBERT: A causal adventure on memetic landscapes. *arXiv*. https://arxiv.org/abs/2410.13488.

[17] Lim, J., Kim, K. (2024, April 14–19). Wav2vec-vc: Voice conversion via hidden representations of wav2vec 2.0. In *Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (p. 10326–10330). Seoul, Republic of Korea.

[18] Bonaccorsi, D. (2023). *Speech-text cross modal learning through self-attention mechanisms* [Doctoral thesis, Politecnico di Torino]. Torino, Italy.

[19] Wang, Y., Xiao, B., Bouferguene, A., Al-Hussein, M., Li, H. (2022). Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning. *Advanced Engineering Informatics, 53*, Article 101699.

[20] Ge, Y., Ge, Y., Liu, X., Wang, J., Wu, J., Shan, Y., Qie, X., Luo, P. (2022). Miles: Visual BERT pre-training with injected language semantics for video-text retrieval. In *Proceedings of the European Conference on Computer Vision* (p. 691–708). Springer.

[21] Nguyen, C. D., Vu-Le, T. A., Nguyen, T., Quan, T., Luu, A. T. (2023, October 29–November 3). Expand BERT representation with visual information via grounded language learning with multimodal partial alignment.

In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 5665–5673). Ottawa, ON, Canada.

[22] Ji, X., Sun, L., Huang, K. (2025). The construction and implementation direction of personalized learning model based on multimodal data fusion in the context of intelligent education. *Cognitive Systems Research, 92*, Article 101379.