ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

A Big Data-Based Model for Analyzing EFL Learners' Language Proficiency in English Language Education

Qi Wang[1,*] *and Bing Li*[2]

[1] Henan University of Animal Husbandry and Economy, Zhengzhou, Henan, 450000, China
[2] Henan University of Animal Husbandry and Economy, Zhengzhou 450046, China.

SUMMARY: *The study combines the ADDIE instructional design model with the SPOC blended EFL English teaching model implementation model designed for the characteristics of university English teaching, and conducts teaching practice based on the model. Firstly, data collection on learning behaviors is carried out, including check-in time, live participation, video learning, etc. as well as data collection on EFL learners' language proficiency. Then Pearson's correlation and multiple regression model are used to construct a model for analyzing the factors influencing the language proficiency of EFL learners, and finally the study is empirically analyzed. The results show that cell phone becomes the main learning mode of learners and they are more willing to accept learning content in video format. The self-assessed mean value of language awareness dimension was the highest at 3.89, followed by language comprehension dimension with a mean value of 3.84, language expression dimension, and language sense dimension with the lowest self-assessed mean value, and the influencing factors on the language proficiency of EFL learners in descending order of influence were: teacher-student language interaction>language attitude>language learning strategy>language knowledge>parental support.*

KEYWORDS: *ADDIE model; multiple regression; university English; language proficiency; instructional design*

1 Introduction

The English non-native language (EFL) learners' language proficiency analysis model of big data is the main embodiment of big data analysis model in the field of education, which is of great significance in English education by using big data and artificial intelligence technology to analyze the English language proficiency of the learners, so as to provide them with personalized proficiency assessment and enhancement strategies [1-4].

In English education, the generation of EFL learners' language proficiency is an extremely complex and slow process [5]. Factors such as learning psychology, cultural background, behavioral patterns, frequency of use, lifestyles, and thinking habits are intertwined and symbiotic, which seriously constrain students' mastery of English language proficiency [6, 7]. Regarding the research on factors affecting EFL learners' language proficiency, literature [8] analyzes the factors affecting EFL learners' language proficiency, discusses student-related factors, teacher-related factors, and context-related factors, and proposes effective measures to enhance language proficiency. Literature [9] aimed to examine the factors affecting the language learning effectiveness of EFL learners, and through an interview with a teacher, the

*wangqi2172@126.com
https://doi.org/10.65102/is2026718
Published: 30 April, 2026

results pointed out that positive intrinsic motivation, social environment, and learning attitudes contribute to the effectiveness of English language learning and vice versa. Literature [10] describes the influence of factors such as motivation, grammar, vocabulary and intensity of instruction on EFL learners' language proficiency, emphasizing that these factors can largely be controlled by the teacher and play an important role in English language teaching and learning. Literature [11] aimed to emphasize the importance of focusing on the factors affecting EFL learners' oral proficiency and based on the literature review, it was stated that learners prioritize and need to pay more attention to the effective teaching of oral skills. Literature [12] analyzed various factors affecting the language proficiency of EFL learners and through proficiency testing it was shown that the length of instruction, the type of school, and the frequency with which learners use English for reading were the main influencing factors. The existence of these problems seriously hinders the improvement of EFL learners' language proficiency, and the introduction of a big data-based model for analyzing EFL learners' language proficiency provides technical support to solve these problems.

By analyzing multi-latitude data such as students' learning habits, learning interests, learning performance, learning behaviors, social relationships, etc., the model can understand students' individualized needs and potential difficulties, thus facilitating teachers to challenge their teaching strategies and provide better teaching services for students [13-16]. For the application of big data language proficiency analysis model and its related technology in realizing personalized teaching of English, literature [17] points out the limitations of traditional teaching methods and emphasizes the importance of using big data technology in EFL learning, which provides personalized teaching by analyzing the students' learning status and behaviors in order to meet the students' differentiated needs and improve the teaching effect. Literature [18] introduced a personalized learning model based on big data in order to improve the learning effect of EFL learners, which uses real-time data processing to dynamically adjust the learning paths and contents, which can meet the personalized learning needs of students and greatly change the traditional language education model. Literature [19] points out the shortcomings of traditional English education methods and emphasizes that the application of big data technology can integrate students' personalized preferences and real-time contextual information to recommend personalized learning resources, which helps to improve students' motivation and learning efficiency.

In addition, it can accurately predict students' future academic performance based on their historical performance and other relevant data, and help teachers develop personalized tutoring plans [20, 21]. In this regard, literature [22] proposed a prediction model of English performance based on big data technology, which can accurately predict students' learning performance based on the analysis of their learning behaviors, historical grades, learning status and other data, and verified its effectiveness through experiments. Literature [23] develops a predictive analytics model based on machine learning, which is able to analyze a large amount of data related to students to improve the accuracy of predicting their academic performance, and verifies it based on experiments. And by utilizing natural language processing, machine learning and other technologies, the model can intelligently assess students' English homework and provide teachers with faster and more accurate homework feedback [24, 25]. Literature [26] pointed out the shortcomings of traditional English homework assessment methods, proposed an automatic grading model based on natural language processing and machine learning, and verified that the model has high performance and has greater potential for application in English homework assessment. Literature [27] points out that there are word spelling and grammar errors in English homework of college students, in order to improve the efficiency of homework assessment, an intelligent correction scheme based on natural language processing technology is proposed, and it is verified that the scheme has a high accuracy rate.

These advantages in English education can help educational institutions better understand students' learning needs and problems, provide personalized educational services, and ultimately improve the overall quality of education.

A blended teaching model of EFL English based on SPOC was constructed on the basis of the ADDIE instructional design model to collect data on learners' learning behavior data and language proficiency. Based on the analysis of the personal information, learning methods and interaction activities of 6,452 EFL learners, a model for analyzing the influencing factors of language ability of FL learners was constructed using Pearson correlation coefficient and multiple regression model. On the basis of this model, a correlation analysis was conducted between the language ability of FL learners and the influencing factors of language ability of each EFL learner. Moreover, using multiple linear regression analysis, the influence magnitudes of each influencing factor on the language ability of EFL learners were determined.

2 SPOC-based EFL hybrid English education model

2.1 Design of SPOC-based blended EFL English teaching model

The ADDIE model is a more mature instructional design model in the West, and it is also a widely used and discussed instructional design model at present.The ADDIE model has different roles in different phases: the analysis phase is to determine the needs, that is, the problems that are to be solved by using teaching and learning, and to form the instructional purposes of the course accordingly; the design phase is to translate the instructional purposes into the objectives of the units, to arrange the unit sequences, and to determine the learning activities of each unit The development stage involves determining the types of learning activities and materials, drafting and piloting learning materials; the implementation stage involves purchasing learning materials and providing assistance and support for learning activities; and the evaluation stage involves maintaining and modifying the content of each stage throughout the model, in addition to implementing student and instructional evaluations.

Based on the ADDIE instructional design model, this study argues that the design process regarding the SPOC blended EFL English teaching model can be similarly categorized into five phases: analysis, design, development, implementation and evaluation, as shown in Figure 1.

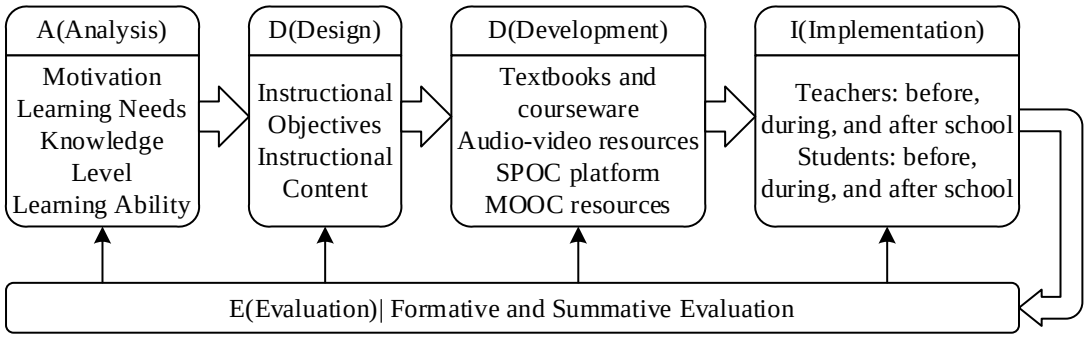

Figure 1: Flow Chart of EFL Teaching Model Design

2.2 Implementation of SPOC-based blended EFL English teaching and learning

Applying the ADDIE instructional design model and combining the characteristics of university English teaching, this study constructed a model for the implementation of a hybrid EFL English teaching model based on SPOC, as shown in Figure 2.

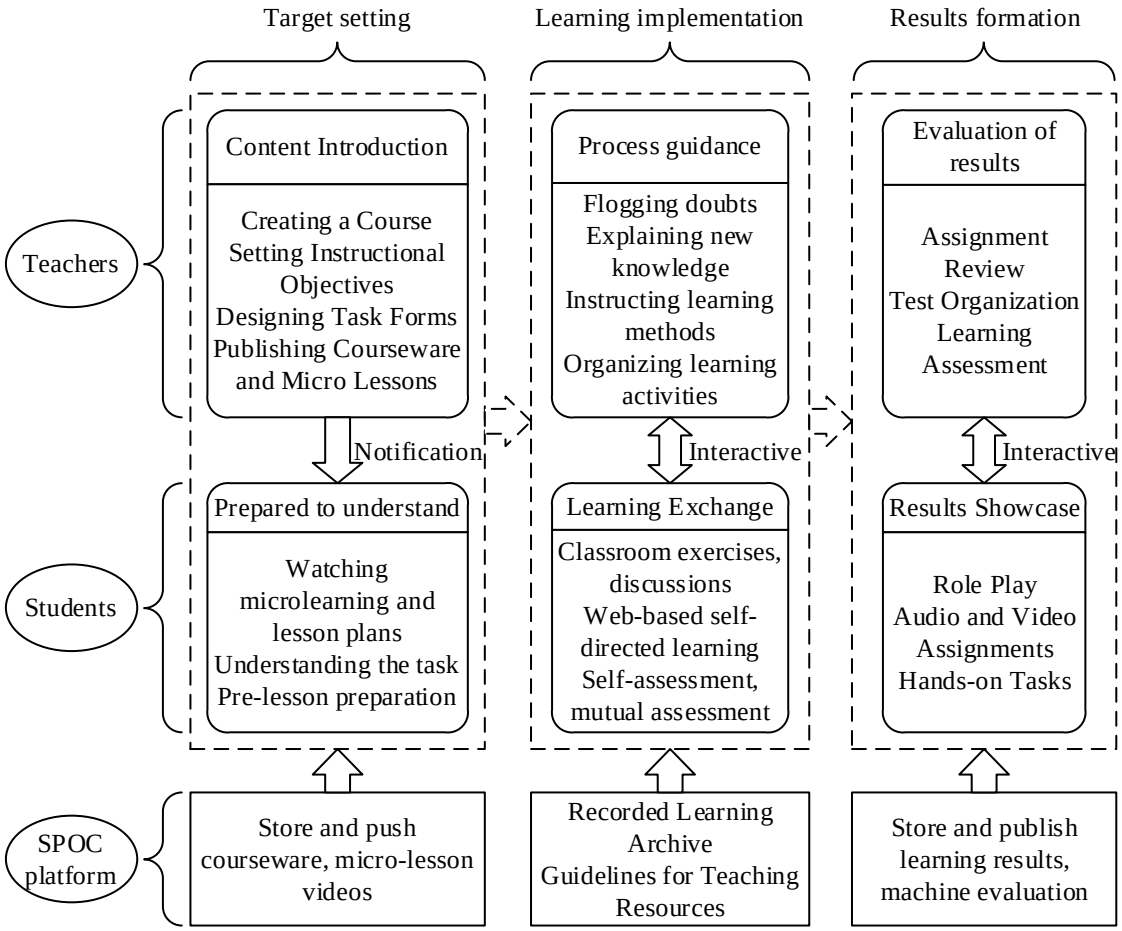

Figure 2: Implementation Model of Hybrid EFL Teaching Mode Based on SPOC

The implementation of the blended English teaching model based on SPOC can be divided into three phases, namely: the goal-setting phase, the learning implementation phase and the results formation phase:

(1) Goal Setting Phase

Before the overall teaching implementation, teachers first select an applicable SPOC platform, establish courses on it, and guide students to log in and register into the corresponding classes. At present, there are a variety of SPOC online classrooms based on cloud technology, and there are also SPOC platforms built on mobile devices. In the era of "Internet +", mobile Internet has been popularized, and if possible, the SPOC platform supporting mobile devices will be the first choice of teachers.

(2) Learning implementation stage

The learning implementation phase includes offline and online classrooms, and the offline and online classrooms are mixed. The form of classroom flipping can be diversified, such as practicing instead of lecturing, student lecture and student evaluation, student debate and teacher evaluation. Regardless of the form, in the offline classroom, teachers can focus more on the explanation and practice of difficult knowledge, answer questions and solve puzzles, guide the learning method, and organize learning activities, while the content of expanding knowledge can be arranged for students to complete on the SPOC platform. Some classroom activities such as testing, quizzing and collective discussion can also rely on the platform, and the SPOC platform becomes a tool for classroom teaching.

(3) Results formation stage

The form of post-course assignments set by teachers includes objective tests, subjective

quizzes, performances and other forms. For the objective test questions, the papers can be assessed through online self-assessment, and the test results can be analyzed, and student-student interactive activities can also be arranged moderately, such as student-assisted teaching, student-student mutual evaluation, and so on. Teachers can grasp the teaching effect in time, so as to revise the teaching program and guide students' learning methods. Students can make use of the rich learning resources on the network and various audio and video production and learning software to carry out independent learning and group learning, and demonstrate the language learning results by combining online audio and video assignments and offline live role-playing, demonstration of results and hands-on tasks.

3 EFL hybrid English language education data mining

3.1 Learning data collection and pre-processing

The data are taken from the learning data of the management background of SPOC courses in University English of a university, and the data of this university is selected because the school has customized and deployed a mode suitable for the learning plan of this school for the actual situation, and the teacher resources are fully utilized from the preparation of the lessons to the lectures, which is of great significance for the research and the subsequent development of this paper. SPOC course data was chosen because the platform is committed to providing scientific course teaching informatization services for university classrooms, and it is an online education platform with a relatively high usage rate by teachers in China at present.

In order to explore the performance of the same students in different semester courses, two semesters of SPOC data were selected for analysis. At the same time, the performance of high school entrance examination scores among freshmen who initially took the course was also considered as a research characteristic, as it reflects the students' enrollment ability to a certain extent.

3.1.1 Check-in time

Check-ins are valid for half an hour after the instructor posts the check-in time, and the instructor sends the check-in through the Learning Access platform prior to the live broadcast.

3.1.2 Teachers' live streaming

Teachers live in the course, whether the students are listening to the whole course, listening to the length of time whether to achieve the length of teaching, whether some students desert and leave in the middle of the course, if the students are required to open the camera, the teacher can learn about the students in the class, but the data can not be recorded. If students do not turn on the camera, neither the data nor the teacher can monitor it, so online teaching also has these problems and defects.

3.1.3 Video learning

The length of video learning, both to see whether the learner in the learning times, drag the progress bar, while setting the task point, the process of video learning, there are test questions, focusing on the learner's mastery of the knowledge of this section as well as the degree of video learning.

3.1.4 Thematic discussions

Whether students actively participate in the thematic discussions delivered by the teacher, and whether they have deep appreciation and insight into the issues involved in the thematic discussions through learning the content.

3.1.5 Operations

Learners in the pre-study can first test, find their own pre-study in the problem, when the teacher live after the end of the test questions, again on the test questions in the difficult problems and questions to answer, test the mastery of the content of this section, the students submit homework, the objective questions can see the results, subjective questions need to be corrected by the teacher before you can see the results.

3.1.6 Classroom performance grades

Grab quizzes during live streaming, group tasks, group chats, and in the classroom the instructor may hand out test questionnaires, all of which have grade points.

$$G = P \times 0.6 + X \times 0.4 \tag{1}$$

Equation (1) is the final grade at the end of the semester, where P is the usual grade, X is the online grade, and G is the student's final course grade, where P is the usual grade given by the instructor during the course of the lecture, which includes the usual grade of the class as well as the final exam.

3.1.7 Data pre-processing

1. College entrance examination scores

Since students come from different provinces, the full score of the college entrance examination varies from province to province, so the processing score of the college entrance examination is needed. Assuming that a student's HKALE score is qs and the full marks of the student's province is ts, the processed HKALE score for the student q is:

$$q = (qs / ts) * 100 \tag{2}$$

2. Mapping course scores to grades

For example, out of 100 points, a score of 60 passes, and a score below 60 fails the course. Above 60 points, every 10 points belongs to a grade and there are 5 grades, A, B, C, D and E. Since the raw full marks of each grade are different, they are divided by percentage.

3. Division of student scores

The scores of the POC platform reflect the students' academic performance on the SPOC platform, and the scores of the offline courses reflect the students' participation in the offline courses and completion of the daily assignments, in addition, the final exam is a way to check the students' mastery of knowledge at the end of the week. Students' scores in this study were calculated as the following four scores: the SPOC platform score, the SPOC platform post-course assignment score, the final exam score, and the college entrance exam score.

3.2 Data collection on language proficiency of EFL learners

This paper uses a self-made EFL learner language ability questionnaire to collect student data. It consists of four sub-dimensions: language comprehension ability, language expression ability, language awareness, and language sense, with a total of 20 items. The Likert five-level positive

scoring scale is adopted, and each item contains five levels from "completely inconsistent" to "completely consistent". The scores are recorded from low to high as 1 to 5 points. The overall alpha coefficient of the questionnaire is 0.885, and the KMO value is 0.892. It is suitable for factor analysis and has good reliability and validity.

3.3 Analytical Model of Factors Influencing Language Proficiency of EFL Learners

3.3.1 Pearson's correlation coefficient

Pearson's correlation coefficient (PCC) can be used to assess the linear correlation between two variables and can be calculated by the following equation:

$$PCC = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}} \tag{3}$$

In this study, x represents the influencing factors and y represents the EFL learners' language proficiency. Pearson's correlation coefficient ranges from -1 to 1. The value tends to 1 (-1) when the positive (negative) correlation between the two variables is stronger, and tends to 0 when there is no significant correlation between the two.

3.3.2 Multiple linear regression

Regression analysis is a statistical process used to assess and explain the relationship between independent and dependent variables, and the linear regression model is one of the most common of the many regression models, and when the number of independent variables in a regression is greater than one, it is referred to as multiple linear regression. In this paper, the multiple linear regression model was used to assess the relationship between EFL learners' language proficiency and the factors influencing each EFL learner's language proficiency, and the results were used to compare and validate the results of the random random forest regression model. The image of the multiple linear regression model is a straight line and the highest order of each independent variable is 1. The multiple linear relationship between the dependent variable $\hat{y}$ (language proficiency of EFL learners) and the independent variable x (influencing factors) can be calculated by the following equation:

$$\hat{y} = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + b \tag{4}$$

where a_i is the regression coefficient of EFL learners' language proficiency, b is the regression constant, and n is the number of influencing factors.

4 Behavioral and linguistic proficiency analysis of EFL learners

4.1 Analysis of EFL Learning Behavior Data

4.1.1 Personal information

Taking the fall semester 2024 of the undergraduate English major College English B course as

an example, there are 6,452 learners participating in the course. The normal distribution of the age of English learners is shown in Figure 3. Through big data analysis, these learners involve all age groups, with the youngest learner being 17 years old and the oldest learner being 59 years old, and the distribution of students at other stages meets the requirements of normal distribution, except for students under eighteen years old who have a more unbalanced age distribution. The average age of over six thousand learners was 26.43 years old, with a median of 26 years old. 5% of the learners were under the age of 18 years old. 31.4% of the learners were distributed between the age groups of 18-25 years old, 22.3% of the learners were distributed between the age groups of 26-30 years old, and 30.1% of the learners were distributed between the age groups of 31-40 years old. Learners over the age of 41 accounted for 8.5% of all participants and those over the age of 51 accounted for only 2.7% of all learners.

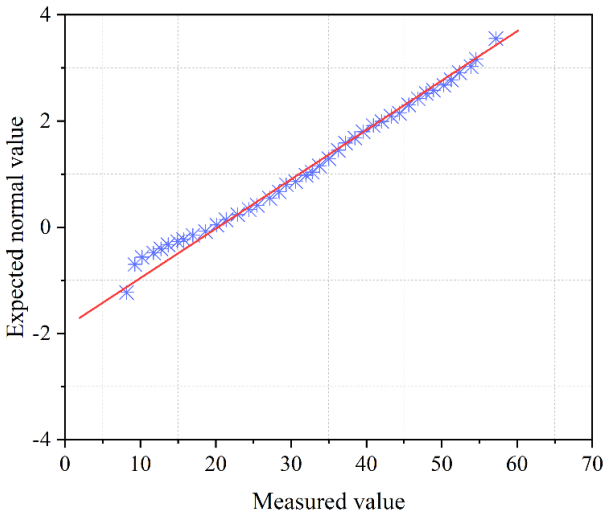

Figure 3: The q-q Diagram of Normal Distribution of Age of Learners of Network English

The common devices used by EFL learners for online learning are shown in Figure 4. The analysis of the Internet devices used by 6452 learners who have participated in the course shows that the most common Internet device used by learners is the cell phone, accounting for nearly 85.74% of the total; 61.76% of the learners use computers to access the Internet; the rate of using tablets for learning is only 8.42%; and there are a very small number of learners (1.58%) who use other devices.

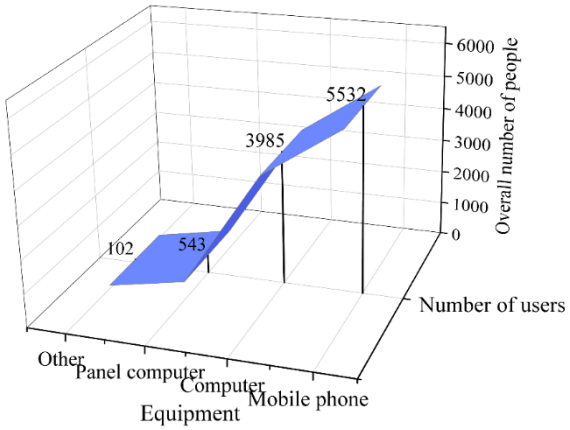

Figure 4: The Situation of the Common Equipment Used by Network English Learners

4.1.2 Learning styles

The analysis of online English learners' learning styles is developed from the big data in terms of time spent participating in online course learning, course text content and video browsing.

(1) Time of Participation in Online Courses

The time distribution of learners' participation in the course is shown in Figure 5. Taking the learners of the undergraduate English major College English B course in the fall semester of 2024 as an example, the analysis of the participation in the course by semester dimension (i.e., month) is used first. The data show that during the whole learning cycle of September-December in the fall of 2024, the time when learners logged into the platform to complete the shape test was concentrated in the three months from September to November, of which November was the peak period for completing the course. The period during and around the holiday break is a low point in the entire learning cycle except for the beginning and end of the semester.

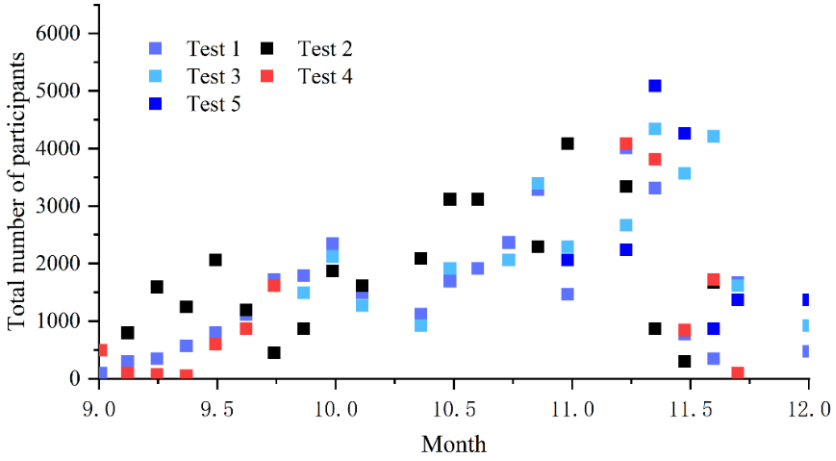

Figure 5: Participants' course attendance distribution by month and day

Analyzing learners' participation in the course from the single-day dimension (i.e., time period) is shown in Figure 6, where learners' participation in the course in the single-day 0:00-24:00 cycle is concentrated in the 9:00-24:00 time period, with the peaks of learning in the 9:00-11:00 and 14:00 -17:00, but there is also a high level of activity in the evening 18:00-24:00 timeframe.

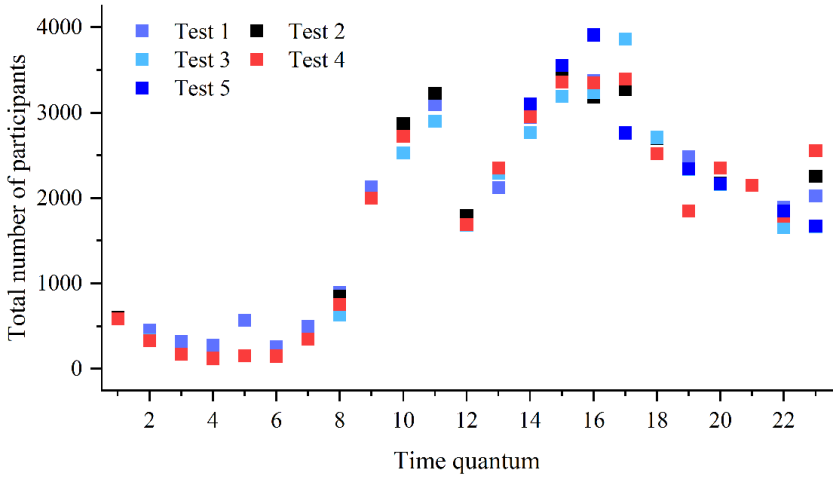

Figure 6: Distribution of learners' time spent in the course

(2) Analysis of Course Text Content and Video Viewing Data

In order to examine the text content and video situation of learners studying English courses, the e-learning data of a total of four semesters of three courses, College English B for undergraduates majoring in English and College English 1 and College English 2 for specialists not majoring in English, were selected as shown in Figure 7. From the perspective of undergraduate English majors, the number of learners browsing videos in Fall 2024 is higher than the number browsing texts, and the number of learners browsing videos in Spring 2025 is even much higher than the number browsing texts, and the difference is even more obvious. Cross-sectional analysis shows that although more people took courses in Fall 2024 than in Spring 2025, however, the number of people who viewed text and videos in Spring 2025 was significantly higher than the number in Fall 2019, and there is an evolving trend. Over time, the number of videos viewed per capita is higher than the number of texts viewed per capita. It can be seen that EFL learners are more comfortable with learning content in the form of videos as opposed to texts such as web pages and PPTs.

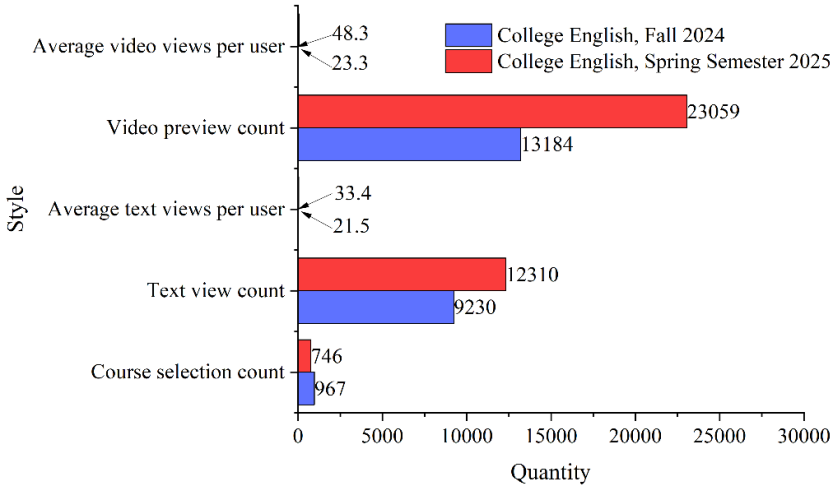

Figure 7: Text and Video Learning Analysis of College English Online Course

4.1.3 Interactive activities

The posting and reposting situation of EFL learners of undergraduate English majors “College English B” and English majors' specialties “College English 1” and “College English 2” in fall 2024 and spring 2025 were selected as samples for data collection to analyze the interaction of learners carrying out course learning, as shown in Figures 8 and 9.

From the data analysis, it can be seen that the number of learners of the undergraduate English major College English B in spring 2025 is slightly lower than the number of people who took the course in fall 2024, but the number of posts is higher than that of the learners in fall 2024, with a per capita number of 15.4 posts, and per capita number of 10.9 posts in fall 2024. However, the difference in postings between Fall 2024 and Spring 2025 was not significant. Similarly, learners of English majors specializing in College English 1 and College English 2 had a higher number of posts in spring 2025 than in fall 2024, with 16.7 posts per capita, compared to 6.2 posts per capita in fall 2024.The difference in posting between fall 2024 and spring 2025 was more significant.

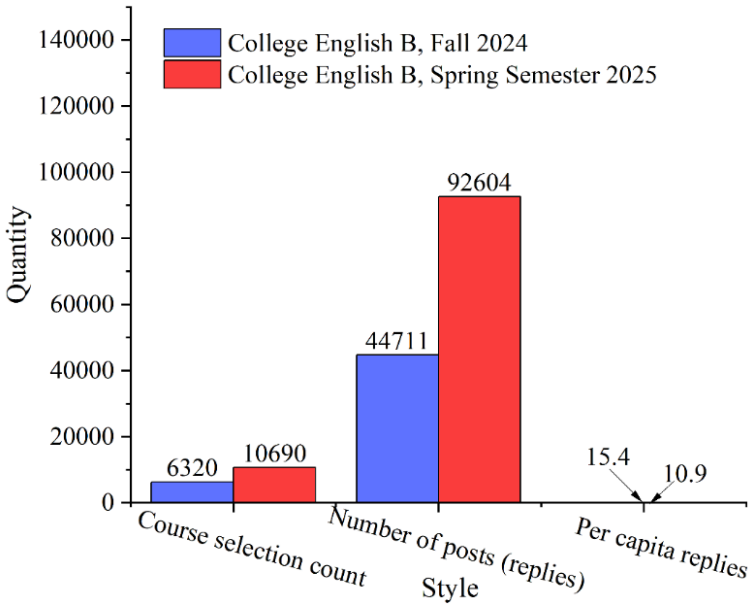

Figure 8: A Comparative Study on the Posting Behavior of English Majors

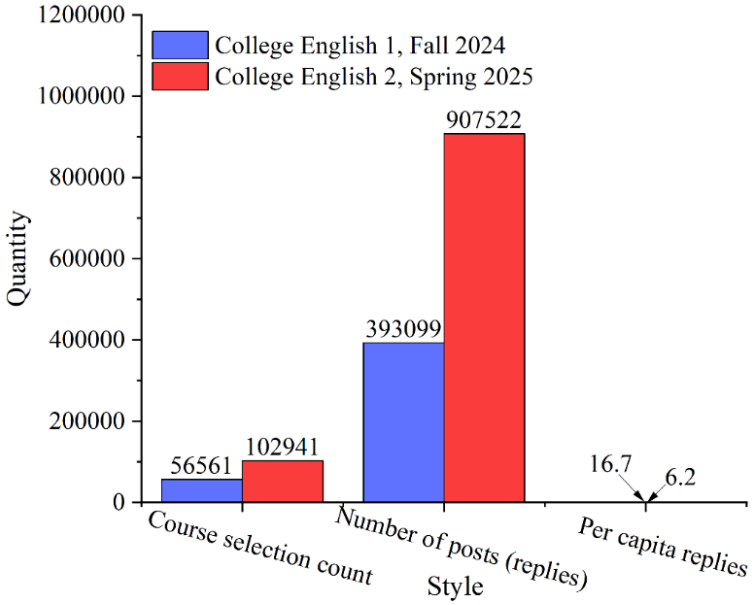

Figure 9: A Comparative Study on the Posting Behavior of College Students Majoring in English

4.2 EFL Learner Competency Analysis

The results of the survey are shown in Table 1. In general, the mean of the overall self-assessment score of EFL learners' language proficiency was 3.73; In terms of the dimensions, the self-assessment mean of EFL learners was 3.84 in the dimension of language comprehension, 3.65 in the dimension of language expression, 3.89 in the dimension of language awareness, and 3.61 in the dimension of language sense. The mean values of the EFL learners' language proficiency dimensions show that the EFL learners performed best in the language awareness dimension, followed by the language comprehension dimension, then the language expression dimension, and finally the language sense dimension.

Table 1: The Overall Situation of Learners' English Language Competence

Dimension	Least value	Crest value	Average value	Standard error	Variance
Language comprehension ability	2.0	6.0	3.84	0.78	0.59
Language competence	2.0	6.0	3.65	0.84	0.68
Language consciousness	2.0	6.0	3.89	0.88	0.74
Language sense	2.0	6.0	3.61	0.95	0.89
Ensemble	-	-	3.73	0.63	0.41

4.3 Analysis of EFL learners' influencing factors

4.3.1 Correlation analysis

In order to investigate the relationship between EFL learners' language proficiency and its influencing factors, correlation analyses were conducted on the dependent variable language proficiency, and the independent variables language knowledge, language attitudes, language learning strategies, teacher-student language interactions and parental support. The degree of correlation between the independent and dependent variables was measured using Pearson's product-difference correlation method as shown in Table 2.

Overall, the correlation coefficients between EFL learners' language proficiency and each of the influencing factors were in the range of 0.464-0.522, and there was a moderate correlation between EFL learners' language proficiency and each of the influencing factors. In terms of the magnitude of the correlation coefficients of each influencing factor, the highest correlation coefficient between teacher-student language interaction and language proficiency is 0.522, followed by language knowledge, language attitude, and language learning strategy, with correlation coefficients of 0.511, 0.508, and 0.481, respectively, and lastly, parental support, with correlation coefficients of 0.464. And the p-values are all less than 0.01, indicating a significant correlation at at the 0.01 level (bilaterally), suggesting that there may be a regression effect between the four influences and proficiency.

Table 2: Correlation Analysis of Factors Affecting English Language Competence

Variable	Language knowledge	Language attitude	Language learning strategy	Teacher-student interaction	Teacher-student interaction	English language proficiency
Language knowledge	1					
Language attitude	0.424**	1				
Language learning strategy	0.255**	0.274**	1			
Teacher-student interaction	0.274**	0.315**	0.387**	1		
Teacher-student interaction	0.311**	0.384**	0.415**	0.292**	1	
English language proficiency	0.511**	0.508**	0.481**	0.522**	0.464**	1

Note: **Significantly correlated at the 0.01 level (two-sided), *Significantly correlated at the 0.05 level (two-sided)

4.3.2 Variable Selection and Model Setting

In order to investigate the magnitude of the influence of each factor on the language proficiency of EFL learners, this study used multiple linear regression equations to deeply analyze the relationship between the language proficiency of EFL learners and the influencing factors. The study set the model as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \quad (5)$$

In conducting the regression analysis, the language proficiency score was scored as the dependent variable (set to Y), and the independent variables were selected as language knowledge (set to X_1), language attitude (set to X_2), language learning strategies (set to X_3), teacher-student language interaction (set to X_4) and parental support (set to X_5). Where $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the partial regression coefficients of each item, which indicates the degree of influence of the independent variable on the dependent variable. β_0 is the constant term, also called the information residual, which represents the non-random portion that cannot be explained by the independent variable but is stable over time. ε is the random error term, which represents the offsetting error that is unavoidable due to the random effects of many factors.

4.3.3 Regression analysis

1. Basic Tests

(1) Coefficient of determination R2 and serial correlation test

Using stepwise regression method, five independent variables, namely, language knowledge, language attitude, language learning strategy, teacher-student language interaction and parental support, are input step by step to form five regression models, and the results are shown in Table 3. Formation of five regression models five models R^2 values become higher and higher with the input of independent variables, explaining more and more about the dependent variable EFL learners' language proficiency. Finally the adjusted R^2 of model five in this study is 0.512, which is a large effect size, indicating that the five independent variables in this study can explain 51.2% of the variance in EFL learners' language proficiency, and the remaining 48.8% may be affected by background variables or other factors, which need to be further explored.

Serial correlation is a test of whether there is a correlation between the various items of the sequence of random error terms of the overall regression model, and serial correlation is usually diagnosed in research by using the Durbin-Watson test (referred to as the D-W test). If the D-W value is closer to 0, it indicates a higher degree of positive serial correlation, and if the D-W value is closer to 4, it indicates a higher degree of negative serial correlation, both of which indicate the presence of pseudo-regression in the model. If the D-W value is close to 2, it indicates that the series are not correlated and there is no pseudo-regression in the model.

Table 3: Test for sequence correlation in regression analysis (DW statistic)

Model	R	R^2	Adjusted R^2	Standard deviation error	Durbin-Watson
1	0.523[a]	0.271	0.265	0.84254	
2	0.619[b]	0.382	0.374	0.78536	
3	0.675[c]	0.455	0.445	0.74093	
4	0.706[d]	0.497	0.483	0.73329	
5	0.731[e]	0.533	0.512	0.69516	1.933

a. Predictor variable:(constant), language attitudes.
b. Predictor variables:(Constant), Language attitudes, Language learning strategies.
c. Predictor variable:(Constant), Language attitudes, Language learning strategies, Language knowledge.
d. Predictor variable:(Constant), Language Attitude, Language Learning Strategies, Language Knowledge, Teacher-Student Language Interaction.
e. Predictor variables: (constant), language attitudes, language learning strategies, language knowledge, teacher-student language interaction, parental support.

(2) Diagnosis of multicollinearity

Whether there is multicollinearity in multiple linear regression is mainly tested by the eigenvalues, conditional indices, tolerance, and variance inflation factor (VIF for short). The study shows that the eigenvalues of all independent variables are greater than 0.01, the conditional indices are less than 30, the tolerances are greater than 0.01, and the VIF values are less than 10, indicating that there is no multicollinearity among the independent variables. The test for multicollinearity is shown in Table 4, and the results all meet the standard requirements, making them suitable for regression analysis.

Table 4: Multiple Collinearity Diagnosis in Regression Analysis

Model	Collinearity statistics			
	Eigenvalue	Condition index	Franchise	VIF
(Constant)	8.433	1.005		
Language knowledge	0.058	12.083	0.715	1.508
Language attitude	0.048	14.271	0.711	1.425
Language learning strategy	0.038	15.083	0.728	1.385
Teacher-student interaction	0.035	16.822	0.672	1.505
Parental Document	0.024	20.615	0.726	1.392

(3) Normal distribution of residuals and variance chi-square test

In the regression analysis, this study determines whether the residuals are normally distributed and whether there is heteroscedasticity by regression standardized residual histogram and regression standardized residual scatter plot, and the test results are shown in Figures 10 and 11. Figure 10 shows that the residuals of this study conform to normal distribution, indicating that the fitting effect is good. From Figure 11, it can be seen that no matter how the predicted values in a specific range change, the corresponding residuals are always near the 0 level line, and the amplitude of the wave remains basically stable, with no obvious signs of heteroscedasticity. It can be determined that there is no significant difference between the standardized residuals and the standardized normal distribution, which indicates that the conditions are met and the regression model is appropriate and feasible.

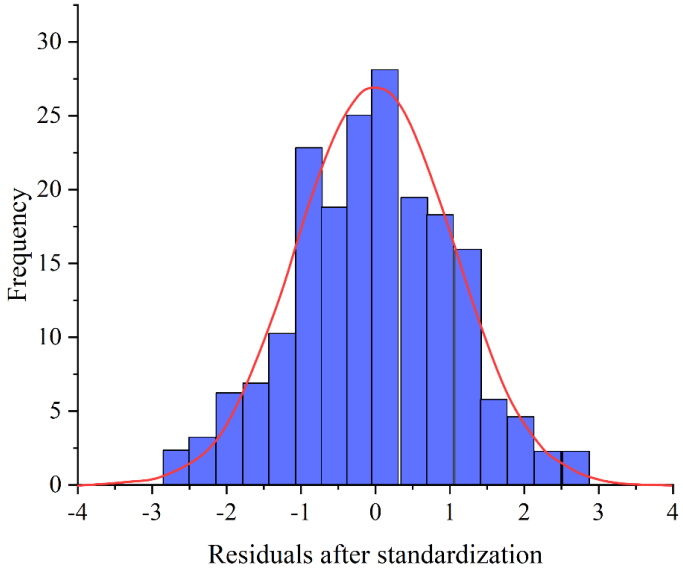

Figure 10: Histogram of regression standardization residual

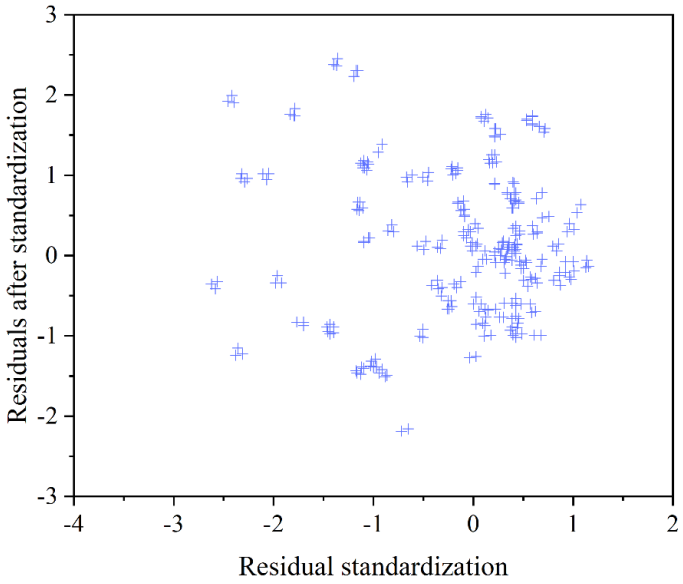

Figure 11: Scatter plot of regression standardization residuals

2. Regression process

The significance test of regression analysis is shown in Table 5. The five independent variables were sequentially entered into the regression model to obtain the unstandardized multiple linear regression equation as:

$$Y = 0.633 + 0.186X_1 + 0.223X_2 + 0.202X_3 + 0.227X_4 + 0.141X_5 \quad (6)$$

The standardized multiple linear regression equation is:

$$Y = 0.176X_1 + 0.214X_2 + 0.195X_3 + 0.224X_4 + 0.135X_5 \quad (7)$$

From the results of the coefficients of the standardized multiple linear regression equations, it is clear that the factors have a positive influence on the language proficiency of EFL learners, namely:

(1) Language knowledge has a more significant positive effect on EFL learners' language proficiency with the fourth highest influence. The coefficient of influence of the factor of language knowledge on the language proficiency of EFL learners is 0.176 and the coefficient of significance is 0.013.

(2) Language attitude has a more significant positive effect on EFL learners' language proficiency with the second rank of influence. The coefficient of influence of the factor of language attitude on EFL learners' language proficiency is 0.214 and the coefficient of significance is 0.003.

Table 5: Equation regression coefficient and significance test

Model	Nonstandardized coefficient		Standard coefficient		
	B	Standard deviation	Beta	T	Conspicuousness
(Constant)	0.633	0.056		12.425	0.039
Teacher-student interaction (X_4)	0.227	0.069	0.224	3.824	0.001
Language attitude (X_2)	0.223	0.074	0.214	3.318	0.003
language learning strategy (X_3)	0.202	0.069	0.195	3.005	0.005
Language knowledge (X_1)	0.186	0.074	0.176	2.952	0.013
Parental support (X_5)	0.141	0.071	0.135	2.004	0.036

5 Conclusion

A SPOC-based blended EFL English teaching model was constructed by applying the five-stage analysis approach of the ADDIE instructional design model by collecting data on learning behaviors (check-in time, live participation, classroom performance, etc.) and related to EFL learners' language proficiency. Pearson's correlation coefficient and multiple linear regression model were used to assess the relationship between EFL learners' language proficiency and the factors influencing EFL learners' language proficiency. In terms of the common devices used by EFL learners for online learning, cell phone was the preferred device, the main study time was concentrated and distributed in the last month at the end of the semester, and the peak of a single day's study was in the time periods of 9:00-11:00 and 14:00-17:00. In terms of learner competence, the mean of the overall self-assessment scores of the language competence of EFL learners was 3.73, which was up to the required level. The results of the correlation analysis and the data from the regression analysis showed that the factors of language knowledge, language attitude, language learning strategies, teacher-student language interaction, and parental support positively contributed to the improvement of EFL learners' language proficiency.

About the Author

Qi Wang was born in Kaifeng, Henan, China, in 1989. She received the B.A. degree in English language study from the University of Henan Province, China, in 2013 and obtained the Ph.D. degree in English language study from Lyceum of the Philippines, in 2022. From 2013 to 2018, she was a research assistant at Henan University of Animal Husbandry and Economy. Since 2019, she has been a lecturer at Henan University of Animal Husbandry and Economy. Her research interests include higher education teaching, English teaching methods and English translation studies.

References

[1] Li, G. (2022). Analysis of matching of corpus input and English proficiency based on the big data neural network model. Advances in Multimedia, 2022(1), 2190873.

[2] Wang, H., Du, Y., & Tsai, S. B. (2021). Evaluation of the Effectiveness Computer-

Assisted Language Teaching by Big Data Analysis. Mathematical Problems in Engineering, 2021(1), 7143815.

[3] Jing, Y., Mingfang, Z., & Yafang, C. (2022). Evaluation model of college English education effect based on big data analysis. Journal of Information & Knowledge Management, 21(03), 2250046.

[4] Shen, Z., Xu, Q., Wang, M., & Xue, Y. (2022). Construction of college English teaching effect evaluation model based on big data analysis. In Proceedings of the 2nd International Conference on New Media Development and Modernized Education.

[5] Liu, H. J. (2015). Use of learning strategies by EFL learners: A study of how it relates to language proficiency and learner autonomy. International Journal of English Linguistics, 5(2), 21.

[6] Bagheri Nevisi, R., & Farhani, A. (2022). Motivational factors affecting Iranian English as a Foreign Language (EFL) learners' learning of English across differing levels of language proficiency. Frontiers in psychology, 13, 869599.

[7] Zarate, M. (2022). Influential Factors Affecting Students' English Proficiency. Journal of Positive School Psychology, 6(7), 3664-3668.

[8] Yen, P. H., Quyen, V. P., & Hien, T. M. (2019). A review of factors influencing learners' gain of English proficiency. CTU Journal of Innovation and Sustainable Development, 11(1), 49-59.

[9] Pariyanto, P., & Pradipta, B. (2019). Factors influencing an EFL learner's proficiency: an English teacher's perspective. Anaphora: Journal of Language, Literary, and Cultural Studies, 2(2), 89-97.

[10] Renandya, W. A. (2013). Essential factors affecting EFL learning outcomes. English teaching, 68(4), 23-41.

[11] Ghafar, Z. N., & Raheem, B. R. (2023). Factors Affecting Speaking Proficiency in English Language Learning: A general overview of the speaking skill. Journal of Social Science (JoSS), 2(6), 507-518.

[12] Santana, J. C., García-Santillán, A., & Escalera-Chávez, M. E. (2017). Variables Affecting Proficiency in English as a Second Language. European Journal of Contemporary Education, 6(1), 138-148.

[13] Ocaña, M., Khosravi, H., & Bakharia, A. (2019, January). Profiling language learners in the big data era. In ASCILITE 2019-Conference Proceedings-36th International Conference of Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education: Personalised Learning. Diverse Goals. One Heart. (pp. 237-245). Singapore University of Social Sciences.

[14] Zang, R., & Wang, L. (2021, April). Personalized teaching model of college English based on big data. In Journal of Physics: Conference Series (Vol. 1852, No. 2, p. 022013). IOP Publishing.

[15] Yan, J., Chen, A., & Wang, H. (2024). An analysis of English learning behavior based on big data and its impact on teaching. Journal of Computational Methods in Science and Engineering, 24(1), 235-251.

[16] Li, X. (2024). A Personalized Teaching System for College English Based on Big Data and Artificial Intelligence. Scalable Computing: Practice and Experience, 25(6), 5477-5485.

[17] Wu, Y. (2023, June). English Learning Analysis and Individualized Teaching Strategies Based on Big Data Technology. In International Conference on Computational Finance and Business Analytics (pp. 421-430). Cham: Springer Nature Switzerland.

[18] Xia, Y., Shin, S. Y., & Shin, K. S. (2024). Designing personalized learning paths for foreign language acquisition using big data: Theoretical and empirical analysis. Applied Sciences, 14(20), 9506.

[19] Guo, Y., & Li, Y. (2025). Exploration of personalized teaching paths in English education based on big data analysis. Journal of Computational Methods in Sciences and Engineering, 25(1), 821-836.

[20] Hou, W. (2022). Analysis of key indicators in English teaching evaluation based on Big Data Model. Scientific Programming, 2022(1), 1231700.

[21] Feng, G., Fan, M., & Chen, Y. (2022). Analysis and prediction of students' academic performance based on educational data mining. IEEE Access, 10, 19558-19571.

[22] Bai, Y., & Guo, J. (2024). Analysis of English Teaching Achievement Prediction Using Big Data. Informatica, 48(9), 75-88.

[23] Flayyih, M. F., & Hassan, T. O. U. T. (2024). Predictive analytics model for students' grade prediction using machine learning. Babylonian Journal of Artificial Intelligence, 2024, 83-101.

[24] Bashir, M. F., Arshad, H., Javed, A. R., Kryvinska, N., & Band, S. S. (2021). Subjective answers evaluation using machine learning and natural language processing. IEEE Access, 9, 158972-158983.

[25] Huang, Z. (2023). An intelligent scoring system for English writing based on artificial intelligence and machine learning. International Journal of System Assurance Engineering and Management, 1-8.

[26] Zhang, D., & Yuan, X. (2022). Intelligent scoring of English composition by machine learning from the perspective of natural language processing. Mathematical Problems in Engineering, 2022(1), 9070272.

[27] Long, J. (2022). A grammatical error correction model for english essay words in colleges using natural language processing. Mobile Information Systems, 2022(1), 1881369.