International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 5758-5767

https://doi.org/10.70517/ijhsa464414

Machine learning algorithms drive research on college entrance examination question assessment

Yuqing Mo[1,*]
[1] Hunan College of Information, Changsha, Hunan, 410200, China
Corresponding authors: (e-mail: lxfmyq@126.com).

Abstract By scientifically assessing the structure, difficulty, knowledge coverage and other factors of the high school examination questions, it can not only provide theoretical support for the proposition, but also promote the fairness and impartiality of the college entrance examination. This paper analyzes and evaluates the questions of the college entrance examination based on the principal component analysis method, and discusses the various factors affecting the quality of the questions of the college entrance examination and their interrelationships. First, data were collected through questionnaire surveys to analyze the 10 factors affecting the questions of the college entrance examination. Subsequently, principal component analysis was applied to downscale these factors and refine three main components, which were the result presentation factor, the test question context factor and the solution process factor. The KMO test and Bartlett's sphericity test on the questionnaire data verified that the data were suitable for principal component analysis, and a principal component with a cumulative contribution rate of 72.437% was finally obtained. The results of the study showed that the difficulty of the test questions, the degree of knowledge synthesis, and the way of presenting information had a significant effect on the overall assessment of the high school test questions.

Index Terms Higher examination questions, Principal component analysis, Test question assessment, Data analysis, Difficulty factor, Information presentation

I. Introduction

As an important means of selecting talents, the college entrance examination has always occupied a pivotal and important position in China's education system [1]. In order to adapt to the development needs of the new era and improve the quality of talent training, China has been actively and steadily promoting the reform of the college entrance examination [2]. Since 2014, the pilot reform of the examination system will be launched, and it is clearly proposed that the college entrance examination will be "regardless of arts and sciences", and it will be comprehensively promoted in 2017. With the advancement of the reform of the college entrance examination, in 2020, two different types of mathematics papers based on the "new college entrance examination" and the "traditional college entrance examination" scheme appeared, among which the new college entrance examination national paper 1 and the new college entrance examination national paper 1 did not distinguish between arts and sciences, with the aim of promoting educational equity [3]-[5]. By the 2023 college entrance examination, 11 provinces and cities have used the test papers under the new college entrance examination reform, and in 2024, 7 more provinces will be added to the new college entrance examination team, until 2025, all provinces in the country will change to the new college entrance examination, and the "traditional college entrance examination" will come to an end [6]-[7]. Based on this background, the college entrance examination question system is also constantly updated, and it needs to have four elements: core values, subject literacy, key abilities, and necessary knowledge. In the traditional analysis of test questions, the quality of test questions is usually analyzed and evaluated by subject experts, and the time period starts at 15 days, and the evaluation focuses on the difficulty of the test questions (there are errors in the evaluation of the difficulty of some subjects), and the analysis of the penetration of subject literacy and core values is too little [8], [9].

With the development of artificial intelligence, machine learning is becoming one of the most dynamic and promising technologies in various fields. It allows computers to make predictions or decisions by learning from data or experience, semantically analyze text, mine semantic associations, and explore semantic ambiguities through models such as BERT models and convolutional neural networks [10], [11]. By constructing a knowledge graph, the relationship between different elements is explored in the vein of visualization [12]. Machine learning provides a path for analyzing and evaluating test questions with its powerful learning ability.

In order to systematically assess the quality of high school examination questions, this study used principal component analysis (PCA) to analyze the various influencing factors of high school examination questions. Principal Component Analysis is able to compress multidimensional data into a few unrelated composite variables, which is important for the handling of a large number of variables in the assessment of test questions. Specifically, the study collects data through questionnaires to identify the key factors affecting the high school examination questions, and applies PCA methods to downsize these factors to distill a few major influencing components. The validity and reliability of the analyzed results are ensured through statistical tests on the data.

II. Principal Component Analysis of Assessment Indicators for Analyzing and Assessing Higher Education Examination Questions

Scientific analysis and assessment of the questions of the college entrance examination is of great significance in realizing the scientific and fair nature of the college entrance examination. Taking this as the starting point of the study, this paper intends to collect data by means of questionnaire survey, so as to further determine the factors affecting the questions of the college entrance examination and explore the interrelationships that exist among these factors by using the principal component analysis method.

II. A. Questionnaire design and implementation

II. A. 1) Questionnaire design

By organizing, analyzing and categorizing the influencing factors mentioned in the existing literature, we have come up with 10 factors affecting the analysis and assessment of high school test questions, which are the degree of synthesis of subject knowledge involved in the test questions, the unfamiliarity of the test question situation, the degree of integrated use of knowledge in the test questions, the interrelationship of information provided by the test questions, the degree of difficulty of the subject knowledge involved in the test questions, the level of thinking of the test questions in examining the students, and the way of presenting test questions' results, The amount of information provided by the test questions, the presentation of information in the test questions and the degree of guessing the answers to the test questions.

II. A. 2) Questionnaire implementation

The questionnaires were distributed on approximately June 22, 2024 This time is the required time of arrival and concentration after the high school entrance examination in the selected school, the number of students concentrating at this point of time is the highest, and the survey is implemented with the help of subject teachers, the probability of validity of the survey data is higher. A total of 258 questionnaires were distributed and 246 were recovered, of which 235 were valid, with a validity rate of 95.53%.

II. B. Principles of Principal Component Analysis

II. B. 1) Fundamentals and basic ideas

1) Basic idea

Principal component analysis is to take the method of mathematical dimensionality reduction, to find a few composite variables to replace the original numerous variables, so that these composite variables can represent as much as possible the amount of information of the original variables, and are not related to each other [13]. This method of statistical analysis, which reduces multiple variables to a few composite variables that are uncorrelated with each other, is called principal component analysis. Principal component analysis is to try to do is the original many variables with a certain degree of correlation, recombined into a new set of unrelated variables to replace the original variables.

2) Basic Principle

According to the definition of the mathematical model of principal component analysis, in order to carry out principal component analysis, it is necessary to derive the principal component coefficients based on the original data, as well as the three conditions of the model, in order to obtain the principal component model. This is the problem that needs to be solved to derive the principal components.

The model is first required to satisfy the following conditions ① F_i, F_j are uncorrelated $(i \neq j, i, j = 1,2,3 \dots p)$ ② The variance of F_i is greater than that of F_2 is greater than that of F_3 ③ The contributions of $F_1, F_2, \dots .. F_p$ add up to equal 1.

According to condition (1) of the principal component mathematical model requires that the principal components are uncorrelated with each other, for this reason the covariance array between the principal components should be a diagonal array. That is, for the principal components:

$$F = AX \tag{1}$$

Its covariance array should be:

$$Var(F)=Var(AX)=(AX)\cdot(AX)'=AXX'A'=\Lambda=\begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_r\end{pmatrix} \tag{2}$$

Let the covariance array of the original data be V , if the original data is normalized then the covariance array is equal to the correlation matrix, i.e., there is:

$$V=R=XX' \tag{3}$$

Then by the principal component mathematical model condition ③ and the nature of orthogonal matrix, if it can satisfy condition ③ it is best to require A to be an orthogonal matrix, that is, to satisfy $AA'=I$, and so the covariance of the original data is obtained by substituting it into the covariance matrix formula of the principal component:

$$Var(F)=AXX'A'=ARA'=\Lambda \tag{4}$$

$$ARA'=\Lambda \quad RA'=A'\Lambda \tag{5}$$

Expanding the above equation yields:

$$\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp}\end{pmatrix}\cdot\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{2p} \\ a_{12} & a_{22} & \cdots & a_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp}\end{pmatrix} \\ \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{2p} \\ a_{12} & a_{22} & \cdots & a_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp}\end{pmatrix}\cdot\begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p\end{pmatrix} \tag{6}$$

Expanding the left and right sides of the equation, by the property of equality of matrices, the equation derived here from the first column only is:

$$\begin{cases}(r_{11}-\lambda_1)a_{11}+r_{12}a_{12}+\cdots+r_{1p}a_{1p}=0 \\ r_{21}a_{11}+(r_{22}-\lambda_1)a_{12}+\cdots+r_{2p}a_{1p}=0 \\ \vdots \\ r_{p1}a_{11}+r_{p2}a_{22}+\cdots+(r_{pp}-\lambda_1)=0\end{cases} \tag{7}$$

In order to obtain the solution of this chi-square equation, the coefficient matrix determinant is required to be 0, i.e.:

$$\begin{vmatrix} r_{11}-\lambda_1 & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22}-\lambda & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp}-\lambda_1\end{vmatrix}=0 \tag{8}$$

Clearly, λ_1 is the eigenvalue of the matrix of correlation coefficients and $a_1=(a_{11},a_{12},\cdots a_{1,p})$ are the corresponding eigenvectors. Similar equations can be obtained based on the second column, the third column, etc., and so are not equations:

$$|R-\lambda I|=0 \tag{9}$$

of p roots, λ_i is the characteristic root of the characteristic equation, and a_j is the component of its eigenvector.

The following is another proof that the variances of the principal components are sequentially decreasing. Let the P eigenroots of the correlation coefficient matrix R be $\lambda_i \geq \lambda_2 \geq \cdots \lambda_p$, and the corresponding eigenvectors a_i :

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} \tag{10}$$

Similarly there is: $Var(F_i) = \lambda_i$, i.e., the variances of the principal components are decreasing in order. And the covariance is:

$$\begin{aligned} Cov(a_i^{'} X^{'}, a_j X) = a_i R a_j &= a_i \left(\sum_{\alpha=1}^{p} \lambda_\alpha a_\alpha a_\alpha^{'} \right) a_j \\ &= \sum_{\alpha=1}^{p} \lambda_\alpha (a_i^{'} a_\alpha)(a_\alpha a_j) = 0, i \neq j \end{aligned} \tag{11}$$

In summary, according to the proof the principal component covariance in principal component analysis should be a diagonal matrix whose elements on the diagonal are exactly equal to the eigenvalues of the correlation matrix of the original data, and the elements of the principal component coefficients matrix A are the eigenvectors corresponding to the eigenvalues of the correlation matrix of the original data. The matrix A is an orthogonal matrix.

Thus, the variables $(x_1, x_2, \cdots x_p)$ are transformed to obtain the new composite variables:

$$\begin{cases} F_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ F_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \cdots \\ F_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{cases} \tag{12}$$

II. B. 2) Calculation steps for principal component analysis

The matrix of sample observations is:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pp} \end{pmatrix} \tag{13}$$

Step 1: Normalize the raw data:

$$x_{xj}^{'} = \frac{x_{ij} - \overline{x}_j}{\sqrt{\text{var}(x_j)}} (i = 1, 2, \cdots, n; j = 1, 2, \cdots p) \tag{14}$$

Among them:

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}, var(x_j) = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2, (j = 1, 2, \cdots p) \tag{15}$$

Step 2: Calculate the sample correlation coefficient matrix:

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix} \tag{16}$$

For convenience, the correlation coefficient of the normalized data is assumed to be X even after normalization of the original data:

$$r_{ij} = \frac{1}{n-1}\sum_{i=1}^{n} x_n x_{ij} \ (j = 1, 2, \cdots p) \tag{17}$$

Step 3: Find the eigenvalues $(\lambda_1, \lambda_2, \dots \lambda_p)$ of the matrix of correlation coefficients R and the corresponding eigenvectors $a_i = (a_{i1}, a_{i2}, \cdots a_{ip}), (j = 1, 2, \cdots p)$.

Step 4: Select the significant principal components and write the principal component expressions.

Principal component analysis can get p principal components, however, because the variance of each principal component is decreasing, the amount of information contained is also decreasing, so the actual analysis, generally do not select p principal components, but according to the cumulative contribution rate of each principal component to select the first k principal components, where the contribution rate refers to the proportion of the variance of a principal component to the total variance, which in practice is the proportion of the eigenvalues to the total of all eigenvalues. The larger the contribution ratio, the stronger the information of the original variable contained in the principal component. The selection of the number of principal components k is mainly based on the cumulative contribution rate of the principal components, i.e., the cumulative contribution rate is generally required to reach 85% or more, so as to ensure that the composite variable can include the vast majority of the information of the original variable.

Step 5: Calculate the principal component scores.

According to the standardized raw data, according to each sample, respectively, substituting into the principal component expression, you can get the new data of each sample under each principal component, that is, the principal component score. The specific form can be as follows:

$$\begin{pmatrix} F_{11} & F_{12} & \cdots & F_{1k} \\ F_{21} & F_{22} & \cdots & F_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ F_{n1} & F_{n2} & \cdots & F_{nk} \end{pmatrix} \tag{18}$$

II. C.Relationship analysis of assessment indicators based on principal component analysis

Since the purpose of principal component analysis is to explore the interrelationships of the influencing factors and the weights of the influences on the questions in the high school exam, we took each factor as a variable and statistically derived the total number of times each factor was selected for each subquestion based on the influencing factors selected by each student for each subquestion. In turn, principal component analysis was performed using SPSS software.

In this paper, the data obtained from the questionnaire was analyzed for principal component analysis using SPSS 17.0. Firstly, KMO test and Bartlett's test of sphericity were done for the influential factors. The results of this KMO and Bartlett's test are specifically shown in Table 1. From the table, it can be seen that the KMO value is 0.705>0.6 and the significance level of Bartlett's test of sphericity is Sig. <0.05, which indicates that this group of data is more suitable for principal component analysis.

Table 1: KMO and Bartlett's test

Kaiser-Meyer-Olkin measure of sampling adequacy		0.705
Bartlett 's sphericity test	Approximate chi-square	262.84
	Df	48
	Sig.	0.003

Next, the fragmentation diagram and the rotated component matrix diagram are made, as shown in Fig. 1. In the fragmentation plot, there are three principal components with eigenvalues greater than 1.

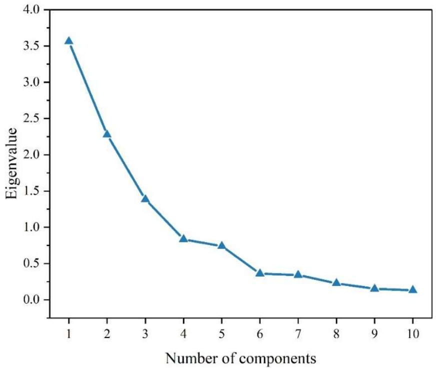

Figure 1: Stone map

The contribution of the rotated factors is specifically shown in Table 2. As can be seen from the table, the cumulative contribution rate of these three principal components reaches 72.437%, which is not easy to reach more than 60% for the data in social and behavioral sciences, indicating that these three principal components reflect most of the information of the original influencing factors, so this analysis selects the first three principal components with eigenvalues greater than 1, a larger contribution rate, and a cumulative contribution rate that meets the requirements.

Table 2: Contribution of the post-rotation factor

Factor number	Eigenvalue	Contribution rate (%)	Cumulative contribution rate (%)
1	2.685	26.854	26.854
2	2.593	25.932	52.786
3	1.965	19.651	72.437

The rotated component matrix is specifically shown in Table 3. If the absolute value of the correlation coefficient is taken as the criterion of whether the correlation coefficient exceeds 0.510 or not, the first principal component mainly summarizes four influencing factors, namely, the comprehensiveness of the subject knowledge involved in the test questions (0.519), the difficulty of the subject knowledge involved in the test questions (0.763), the guessing degree of the answers to the test questions (-0.85), and the way of presenting the results of the test questions (0.912), which can be named the result presentation factor as they are related to the questions to be answered and the presentation of question results, they can be named as result presentation factors. In the rotation matrix, the correlation coefficient of the degree of guessing of test answers is negative, which means that the degree of guessing of test answers is negatively correlated with the degree of difficulty of test questions. The second principal component mainly includes four influencing factors: the presentation of test question information (0.548), the amount of information provided by the test question (0.915), the interrelationship of the information provided by the test question (0.798), and the unfamiliarity of the test question's context (0.83), which are all related to the test question's context, and can be named as the test question context factor; the third principal component mainly summarizes the degree of the comprehensive use of knowledge in the test question (0.611). The third principal component mainly summarizes the two influencing factors of the degree of utilization of knowledge of the test questions (0.611) and the level of thinking of the test questions to test the students (0.883), which are related to the process of solving the test questions and can be named as the factor of the process of solving the test questions.

Table 3: Rotating component matrix

-	Component		
	1	2	3
Presentation of test information	-0.017	0.548	0.503
How much information the test questions provide	-0.15	0.915	0.062
Relationship between the information provided by the test questions	0.227	0.798	0.37
Strangeness of the test situation	0.127	0.83	-0.305
Test questions involve the comprehensiveness of subject knowledge	0.519	0.068	0.398
Test questions involve the difficulty of subject knowledge	0.763	0.082	0.245
Degree of comprehensive application of test knowledge	0.51	0.38	0.611
Test questions examine students ' thinking level	0.035	-0.073	0.883
Guessing degree of the answer to the test question	-0.85	0.014	0.12
Presentation of test results	0.912	-0.016	0.03

III. A machine learning-based model for analyzing and evaluating high school exam questions

In the previous chapter, this paper used the principal component analysis method to clarify the relationship between the factors affecting the analysis and assessment of college entrance examination questions, and in this chapter, we will take the principle of multiple linear regression model as the basis, choose the linear regression method of machine learning to construct a model for the analysis and assessment of college entrance examination questions, and use the actual data of the part of the questions of the Jiangsu volume of the college entrance examination in chemistry for the calibration, to explore the value of the model's application and the significance of its promotion [14].

III. A. Principles of multiple linear regression modeling

Multiple linear regression analysis is a statistical method used to evaluate the relationship between a dependent variable and multiple independent variables. In addition to meeting the conditions of univariate linear regression, multiple linear regression also needs to meet the conditions of multiple independent variables without multicollinearity. Multiple linear regression needs to meet the following conditions: the independent variable and the dependent variable are theoretically causally related; the dependent variable is a continuous variable; there is a linear relationship between the respective variable and the dependent variable; the residuals need to meet the conditions of normality, independence, and chi-square; and multiple independent variables do not have Multicollinearity. Among them, linearity, normality, independence, and variance chi-square are the four basic prerequisites for linear regression analysis.

For a sample i with n features, the multivariate linear regression result is shown in equation (19):

$$y = w_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_n x_{in} \tag{19}$$

In this expression, w is collectively referred to as the parameters of the model, where w_0 is referred to as the intercept and $w_0 \sim w_{\mathrm{n}}$ are referred to as the regression coefficients. Where y is the target variable and $x_{i1} \sim x_{in}$ are the different characteristics on the sample i.

If we consider that there are m samples, and y is a column vector of regression results containing all the samples of m, we can represent this equation using a matrix, where w can be regarded as a column matrix with structure [1, n] and X is a feature matrix with structure [m, n], then the target variable Y is shown in equation (20):

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \cdots \\ y_m \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & x_{13} & \cdots & x_{13} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{23} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{m3} \end{bmatrix} \times \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \cdots \\ w_m \end{bmatrix} \tag{20}$$

The task of linear regression is to construct a measurement function to map the linear relationship between the input feature matrix X and the objective function value Y. The core of constructing a measurement model is to find the parameter vector of the model w. In multiple linear regression, the loss function is defined as shown in equation (21):

$$Loss\ funtion = \sum_{i=1}^{m}(y_i - y)^2 = \sum_{i=1}^{m}(y_i - X_i w)^2 \tag{21}$$

where y_i is the true label corresponding to the sample i and $y_i = X_i w$ is the measured value of the sample i under a set of parameters w. First, this loss function represents the squared result of the L2 paradigm of $y - y_i$, which is essentially the Euclidean distance, i.e., the sum of squares of each point on the two vectors corresponding to the sum of squares of the two vectors subtracted from each other and then squared, and here, the loss function realizes the sum of squares of each point on the vectors corresponding to the sum of squares of the two vectors subtracted from each other, and not squared, and so the loss function is the L2 paradigm of the squared Euclidean distance Result. Under this squared result, y and y_i are the true label and the measured value, respectively, i.e., this loss function is calculating the distance between the true label and the measured value. Therefore, we consider that this loss function measures the difference between the measured results and the true labels of our constructed model, and the model expects that the smaller the difference between the measured results and the true values, the better, so our solution objective can be transformed into equation (22):

$$\min \| y - Xw \|_2^2 \tag{22}$$

where the 2 in the lower right corner indicates the L2 paradigm of the vector $y - Xw$, which is what the loss function represents. Squaring on the L2 paradigm is the loss function, and this equation is often referred to as the sum of squares of the errors or the sum of squares of the residuals.

III. B. Model construction for analyzing and evaluating high school exam questions

III. B. 1) Impact factor assignment

Eight secondary school teachers with rich teaching experience and experience in proposing questions for the college entrance examination were selected to independently assign values to the factors affecting the analysis and evaluation of college entrance examination questions, and the consistency analysis of the teachers' assignment results yielded a Kendall's harmony coefficient of 0.825, which indicated that the assignments of the eight teachers were consistent at the 0.01 level.

III. B. 2) Modeling

The biggest advantage of machine learning is to transform a large amount of unorganized data into useful information [15]. Based on machine learning preferred regression method modeling, multiple linear regression method was selected for modeling, that is, to establish a linear relationship between multiple independent variables (factors influencing the difficulty of the test questions) and the dependent variable (difficulty of the test questions), so that the constructed model of the difficulty of the test questions is not only relatively stable, but also has a good interpretability.

In the process of regression modeling, 70% of all the data are randomly selected as the training set to establish the linear regression equation, and then the remaining 30% of the data are used as the test set for model validation to obtain the mean of the error of the difficulty coefficient and the standard deviation of the error to test the accuracy of the constructed model. In accordance with the principle of machine learning to minimize chance errors, the models with the largest and smallest test errors are discarded, and the model with higher error stability and which can be given an explanatory meaning is selected. According to the actual meaning of the parameters in the constructed model, the coefficients reflect the weight of the variables on the effect of difficulty, thus removing models with negative coefficients of the variables.

The difficulty of all test questions was predicted using the model and the predicted and measured values were compared and the results are shown in Figure 2. The results show that the model predicted values of test question difficulty have a good fit with the measured values.

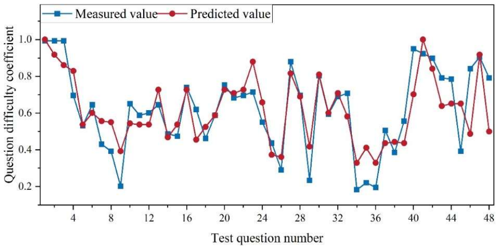

Figure 2: Model predicted value and measured value of the difficulty of testing item

III. C. Model test for analyzing and evaluating high school exam questions

Whether the high school examination question analysis and assessment model constructed using machine learning in this paper is generalizable, whether it can be applied to the actual examination, and whether it can realize a more accurate prediction of the difficulty of the test questions, it needs to be tested with real high school academic level examination questions. The real test data of twelve non-selective questions in the Jiangsu paper of chemistry of the gaokao from 2022 to 2024 are selected for the model test. The model predicted and measured values of the difficulty of the high school examination questions are specifically shown in Figure 3. The test results show that the data predicted using the model have a good fit with the measured data, and the constructed model has certain application value and promotion significance.

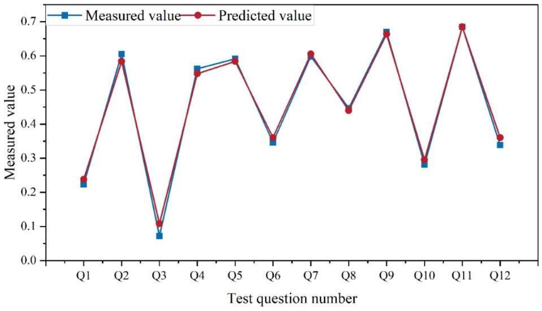

Figure 3: Model predicted value and the measured value of the difficulty

IV. Conclusion

In the analysis and assessment of high school examination questions, the principal component analysis method can effectively refine the main factors affecting the quality of the questions. According to the analysis results, it is found that factors such as the comprehensive difficulty of test questions, knowledge coverage, and the way of presenting information are of great significance to the influence of high school examination questions. Specifically, the difficulty of test questions involving subject knowledge is negatively correlated with the degree of answer guessing, while the way of presenting test results is closely related to the difficulty. The comprehensive analysis of the three main components can help proposers better understand and control the quality of the high school test questions. In the process of model validation in this paper, there is a good fit between the experimental data and the predicted results, indicating that the model has certain practical value and promotion significance. In the future, further research can combine more dimensions of data to improve the prediction accuracy of the model, thus providing stronger support for the scientific and standardization of the college entrance examination.

References

[1] Yao, Y., Zhang, Z., Cui, H., Ren, T., & Xiao, J. (2019). The influence of student abilities and high school on student growth: A case study of chinese national college entrance exam. IEEE Access, 7, 148254-148264.

[2] Jiang, Q., & Guo, X. (2020, February). Research on the reform of Chinese college entrance examination system. In Third International Conference on Social Science, Public Health and Education (SSPHE 2019) (pp. 107-111). Atlantis Press.

[3] Zhu, P., Li, M., & Zhu, Z. (2024). Diversification of subject combinations in the National College Entrance Examination and educational reforms in senior secondary schools: findings from China's policies on college admissions. Asia Pacific Education Review, 1-12.

[4] Jian, L. I. (2020). Exploring new college entrance examination ("xin gao kao") policy in china: National values and regional practices. Beijing International Review of Education, 2(3), 466-471.

[5] Huang, S. (2025). Analysis of the Influence of College Entrance Examination System Reform on Educational Equity. Journal of Social Science Humanities and Literature, 8(1), 1-7.

[6] Cheng, J., & Li, B. (2024). Study on the Factors Influencing the Satisfaction of Students in China's New College Entrance Examination Reform. Journal of Roi Kaensarn Academi, 9(3), 543-556.

[7] Mengna, Z., & Chengwu, R. (2024). Investigation and research on comprehensive quality assessment in the reform of new college entrance examination. South African Journal of Education, 44(1), S1-S18.

[8] Cao, C., & Li, L. (2025). PROBABILITYAND STATISTICS QUESTIONS OF COLLEGE ENTRANCE EXAMINATION BASED ON MATHEMATICS CORE LITERACY--A CASE STUDY OF COLLEGE ENTRANCE EXAMINATION MATHEMATICS IN RECENT FIVE YEARS (2019-2023) ACCOMPLISHMENT. Educational Research and Human Development, 28.

[9] Han, C., & Xiang, J. (2025). Alignment Analysis Between China College Entrance Examination Physics Test and Curriculum Standard Based on E-SEC Model. International Journal of Science and Mathematics Education, 23(1), 215-234.

[10] Kaneda, R., Okada, M., & Mori, N. (2021, July). Estimating Semantic Relationships between Sentences Using Word Embedding with BERT. In 2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI) (pp. 53-58). IEEE.

[11] An, T., Shui, X., & Gao, H. (2022, August). Deep learning based webshell detection coping with long text and lexical ambiguity. In International Conference on Information and Communications Security (pp. 438-457). Cham: Springer International Publishing.

[12] Dessì, D., Osborne, F., Recupero, D. R., Buscaldi, D., & Motta, E. (2021). Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. Future Generation Computer Systems, 116, 253-264.

[13] Anahita Nodehi, Mousa Golalizadeh, Mehdi Maadooliat & Claudio Agostinelli. (2025). Torus Probabilistic Principal Component Analysis. Journal of Classification, (prepublish),1-2.

[14] Abdelghani Benghanem, Olivier Valentin, Philippe Aubert Gauthier & Alain Berry. (2024). Objective quantification of sound sensory attributes in side-by-side vehicles using multiple linear regression models. Frontiers in Acoustics, 2, 1477395-1477395.

[15] Oo Bee Lan, Nguyen Anh Tuan, Ahn Yonghan & Lim Benson Teck Heng. (2025). Predicting the number of bidders in construction competitive bidding using explainable machine learning models. Construction Innovation, 25(7), 158-188.