Fuzzy Cluster Analysis on Influencing Factors of College Student Scores

Jia Wen[1], Xiaochong Wei[2*], Haipeng Liu[2], Yiyuan Rong[2]

[1] School of Foreign Languages, Yibin University, Yibin 644000, China
[2] School of Management, Minzu University of China, Beijing 100081, China

Corresponding Author Email: wxcyyy1220@163.com

ABSTRACT

In the age of the Internet, the learning environment is increasingly diversified. It is of great importance to explore the factors that truly affect the college student scores. Focusing on 13 factors that potentially influence college student scores, this paper carries out a questionnaire survey on students of different grades from different colleges, conducts fuzzy processing of the collected data, randomly selects the processed data for initialization of attribute values. Then, the initialized data were subject to principal component analysis (PCA), fuzzy cluster analysis (FCA), and analysis of variance (ANOVA). Through the analysis, six factors were identified as the key factors affecting college student scores: family factor, Exam factor, exchange factor, learner factor, classmate factor, and campus factor. On this basis, the authors called for the concerted efforts from the school, teachers, and students for improving the teaching quality in colleges.

1. INTRODUCTION

Apart from the growing number of college students, the popularization of education has brought various problems. For example, the settings of subjects and specialties are suboptimal, the capability of scientific research is insufficient, and the research results are not fully applied in practice. These problems have a direct or an indirect impact on the overall quality, learning ability, and research ability of college students.

The above problems are magnified by the proliferation of the Internet and the lax attitude among many college students. The overall quality and abilities of college students can be intuitively reflected by their test scores, which are affected by various factors. To enhance their overall quality and abilities, the key is to identify these factors and their impact mechanisms, and make pertinent optimizations.

To date, college student scores have been compared systematically at home and abroad. The existing studies mainly emphasize on the following aspects: learning attitude and scores; student psychology; major preference and scores [1]; influence of family environment on scores; effects of intellectual and non-intellectual factors on student scores [2].

Many scholars have explored the influence of various factors over college student scores [3]. For instance, Hazrati-Viari et al., Nechita, et al. [4, 5] discussed how the various characters and personalities of students affect their scores through regression analysis. Yigermal et al. [6] performed correlation and regression analyses to reveal the influence of multiple factors (e.g. student origin, gender, and National College Entrance Examination (NCEE) results) on college student scores. Musah et al. [7] disclosed the influence of classroom justice and learning style on college student scores through regression analysis, significance test, and confirmatory factor analysis. Olango [8] established an eight-factor model of learning purpose, learning attitude, etc. and a three-factor model of mathematics self-efficacy, mathematics anxiety, etc., and conducted correlation analysis to explore the impact of the eleven factors on college student scores.

However, the above studies generally presuppose one or several influencing factors, before analyzing whether and how much each of these factors influence colleges student scores, using various empirical methods. Few of them have extracted the factors affecting scores in an objective manner. By factor analysis and analysis of variance (ANOVA), Zhu et al. [9] extracted four main influencing factors of scores from the presupposed factors; But the presupposed factors merely cover three aspects: learning style, teaching style, and exam style.

Drawing on the relevant literature, the college student scores are influenced by various factors in three dimensions, namely, individual, school, and family [10]. On this basis, this paper carries out fuzzy cluster analysis (FCA) on the influencing factors of college student scores. The remainder of this paper is organized as follows:

Section 2, the core of this research, extracts 13 main factors from presupposed influencing factors through fuzzy mathematics, FCA, and principal component analysis (PCA). The mathematical models and data analysis principles were explained in details. The NCEE results and scores of undergraduates in different grades from different colleges were preprocessed into datasets, from which the classification knowledge of factors was obtained by cluster analysis, a data mining technique. Firstly, different types of original data were subject to fuzzy clustering [11], and different membership functions were constructed to initialize the data. Next, a fuzzy matrix was designed based on the Euclidean distance formula. Further, the uncertain data were described mathematically through FCA, laying a solid basis for the classification of influencing factors [12].

Section 3 combines PCA, cluster analysis, and factor analysis to reduce the dimensionality of multiple influencing factors in our problem. The combined strategy can decipher

the meaning of principal components, and eliminate the mutual influence between variables. Firstly, the 13 presupposed factors were processed by clustering in R. Then, PCA was performed on each factor to extract the principal components. The different principal components were merged into the main factors affecting college student scores. After that, the cores and NCEE results were treated by multiple linear regression (MLR) to disclose the impact mechanisms of the main factors.

Section 4 determines six factor groups for ANOVA. Through ANOVA, the leading influencing factor on college student scores was identified in each factor group. In this way, the main purpose of this research was achieved: finding out the main factors affecting college student scores.

Section 5 summarizes the research findings, and puts forward effective measures to improve undergraduate scores in colleges.

2. FCA ON INFLUENCING FACTORS

Student scores are affected by a massive number of factors. The massive data contain lots of valuable information. Cluster analysis, which aims to allocate data objects with similar properties and features to the same cluster, can effectively distinguish the key influencing factors, facilitating the design of pertinent measures to improve student scores.

The data on different students involve many fuzzy concepts (e.g. attention from parents, learning interest, frequency of independent completion of homework, and dormitory atmosphere) that cannot be defined or classified by the set theory in classic mathematics. FCA can provide realistic mathematical description of these uncertain data, and mine out the factors affecting student scores from them [13].

2.1 Fuzzy processing of original data

The original data on the influencing factors of student scores were divided into Boolean data, numeric data, generic data, and null data. The four kinds of data were initialized by membership function.

2.1.1 Membership function of Boolean data

Boolean attributes are relatively simple. In this analysis, only two factors exist as Boolean data: "Preview before class?" and "Participation in clubs and student union?".

Let U be the entire data domain, n be the total number of data in U, and N is the number of yes or no. Then, the membership function of Boolean data can be defined as [14]:

$$u(a) = \begin{cases} \dfrac{N(0)}{n} & a=0 \\ \dfrac{N(1)}{n} & a=1 \end{cases} \quad (1)$$

2.1.2 Membership function of numeric data

Many factors exist as numeric data, such as monthly number of exchanges with tutor, weekly number of self-studies, and mean exchange time with tutor. The numerical attribute values can be classified, putting the same attribute values into the

same class [15].

Let U be the entire data domain, n be the total number of data in U, I be the total number of classes, C_i be the i-th class, and $N(C_i)$ be the number of attribute values in class C_i [16]. Then, membership function of numeric data can be defined as:

$$u(C_i) = \frac{N(C_i)}{n} \quad (2)$$

2.1.3 Membership function of generic data

Generic attribute values are classification attributes, e.g. education of parents, attention from parents, learning interest, learning satisfaction, and frequency of independent completion of homework. The value of each attribute is the common value of a class out of a limited number of classes. Once the same attribute values are allocated to the same class, the membership function will focus on the proportion of each class of attribute values in the total set of classes [17].

Let U be the entire data domain, n be the total number of data in U, J be the number of attribute classes, C_j be the j-th class, and $N(C_i)$ be the number of attribute values in class C_j. Then, the membership function of generic data can be defined as [18]:

$$u(C_j) = \frac{N(C_j)}{n} \quad (3)$$

2.1.4 Membership function of null data

Each null data corresponds to the features of its attribute value. Null values may appear in all the previous three types of data. If the ratio of the number of nulls to the number of total elements in an attribute value surpasses the preset threshold Z_0, then the attribute will not be considered in cluster analysis; if the ratio is below the threshold, the attribute will be classified into three levels (high, medium, and low), corresponding to the membership levels (high, medium, and low).

Let C_{ij} be the value of the j-th attribute of the i-th element; r_0 be the said ratio; h_0 is the high-level threshold; l_0 is the low-level threshold. Then, the membership function of null data can be defined as:

$$u(C_{ij}) = \begin{cases} min, & r_0 \leq l_0 \\ mid, & l_0 < r_0 < h_0 \\ max, & h_0 \geq r_0 \end{cases} \quad (4)$$

2.1.5 Fuzzy processing of the data on influencing factors

The fuzzy processing of the data on influencing factors of student scores was explained through an example. For convenience, 14 attributes were selected as classification attributes. To diversify the original data, an online questionnaire survey was conducted among students from different colleges, who learn on different online education platforms. A total of 248 questionnaires were returned, among which 235 were valid. Due to the sheer volume of data, this paper only presents the results on six factors of the first 50 respondents. But the subsequent analysis still deals with 12 factors of 235 respondents (Table 1).

Table 1. The data on influencing factors of student scores

Serial number	Education of parents	Attention from parents	Preview before class?	Weekly number of self-studies	Learning interest	Mean time of exchange with tutor	...
1	Bachelor	Strong	No	6	Strong	2	
2	Junior high school graduate	Moderate	Yes	5	Neutral	1.5	
3	Primary school graduate	Weak	Yes	0	Strong	1.5	
4	Doctor	Strong	No	6	Neutral	2	
5	Junior high school graduate	Moderate	Yes	4	Weak	1	
6	Bachelor	Strong	No	7	Neutral	3	
7	Primary school graduate	Weak	No	6	Neutral	1.5	
9	Senior high school graduate	Slight	No	5	Neutral	3	
10	Senior high school graduate	Slight	Yes	4	Neutral	1	
11	Junior high school graduate	Moderate	Yes	6	Weak	1	
12	Bachelor	Strong	No	8	Neutral	3	
13	Junior high school graduate	Moderate	No	4	Weak	2.5	
14	Junior high school graduate	Moderate	No	6	Strong	2	
…	...	...	...	...	...	...	…

2.1.6 Initialization of attribute values

(1) Boolean attribute values

1) Preview before class?

$$u(Yes)=\frac{60}{235}=0.2553 \quad u(No)=\frac{165}{235}=0.7021$$

2) Participation in clubs and student union?

$$u(Yes)=\frac{146}{235}=0.6213 \quad u(No)=\frac{83}{235}=0.3532$$

(2) Generic attribute values

1) Education of parents

$$u(Primary\ school\ graduate)=\frac{45}{235}=0.1915$$

$$u(Junior\ high\ school\ graduate)=\frac{54}{235}=0.2298$$

$$u(Senior\ high\ school\ graduate)=\frac{56}{235}=0.2383$$

$$u(Bachelor)=\frac{36}{235}=0.1532 \quad u(Master)=\frac{22}{235}=0.0936$$

$$u(Doctor)=\frac{20}{235}=0.0851$$

2) Attention from parents

$$u(Strong)=\frac{66}{235}=0.2809 \quad u(Moderate)=\frac{54}{235}=0.2298$$

$$u(Slight)=\frac{57}{235}=0.2426 \quad u(Weak)=\frac{45}{235}=0.1915$$

3) Learning interest

$$u(Strong)=\frac{50}{235}=0.2127 \quad u(Neutral)=\frac{84}{235}=0.3574$$

$$u(Weak)=\frac{83}{235}=0.3532$$

4) Weekly nonattendance

$$u(Rare)=\frac{86}{235}=0.3660 \quad u(Never)=\frac{88}{235}=0.3745$$

$$u(Occasional)=\frac{48}{235}=0.2043$$

5) Learning satisfaction

$$u(Strong)=\frac{50}{235}=0.2128 \quad u(Slight)=\frac{125}{235}=0.5319$$

$$u(Weak)=\frac{63}{235}=0.2681$$

6) Frequency of independent completion of homework

$$u(Strongly\ high)=\frac{56}{235}=0.2383$$

$$u(Slightly\ low)=\frac{61}{235}=0.2596$$

$$u(Slightly\ high)=\frac{63}{235}=0.2681$$

$$u(Strongly\ low)=\frac{41}{235}=0.1475$$

7) Influence of roommates

$$u(Positive\ influence)=\frac{50}{235}=0.6383$$

$$u(Negative\ influence)=\frac{47}{235}=0.2$$

$$u(No\ influence)=\frac{131}{235}=0.5574$$

(3) Numeric attribute values

1) Monthly number of exchanges with tutor can be divided into the following intervals depending on the attribute values:

$$d_1:[0, 3];\ d_2:[4,8];\ d_3:[9,12]$$

$$u(d_1)=\frac{118}{235}=0.5021\quad u(d_2)=\frac{66}{235}=0.2808\quad u(d_3)=\frac{45}{235}=0.2255$$

2) Weekly number of self-studies can be divided into the following intervals depending on the attribute values:

$$d_1:[0, 2];\ d_2:[3,5];\ d_3:[6,7]$$

$$u(d_1)=\frac{58}{235}=0.2468\quad u(d_2)=\frac{125}{235}=0.5319\quad u(d_3)=\frac{54}{235}=0.2298$$

3) Mean time of exchange with tutor can be divided into the following intervals depending on the attribute values:

$$d_1:[0, 1];\ d_2:[1,1.8];\ d_3:[1.8,2.5]$$

$$u(d_1)=\frac{66}{235}=0.2809\quad u(d_2)=\frac{93}{235}=0.3957\quad u(d_3)=\frac{67}{235}=0.2851$$

4) NCEE results can be divided into the following intervals depending on the attribute values:

$$d_1:[400, 500];\ d_2:[500,600];\ d_3:[600,750]$$

$$u(d_1)=\frac{33}{235}=0.1404\quad u(d_2)=\frac{99}{235}=0.4596\quad u(d_3)=\frac{92}{235}=0.3915$$

(4) Null attribute values

There is no null attribute value in the original data.

2.1.7 Data initialization

The initial data on influencing factors of student scores are shown in Table 2.

2.2 Clustering of initial data

The initial data were clustered by fuzzy matrix. Let U be the universe (Table 3) containing $|U|$ elements. Then, the clustering was implemented in the following steps.

(1) Establishing the fuzzy similarity matrix R of the universe U

Let $|U|$ be the order of R. The elements r_{ij} of matrix R can be calculated by the Euclidean distance formula:

$$r_{ij}=\begin{cases}1 & i=j \\ \sqrt{\dfrac{1}{m}\sum_{k=1}^{m}(S_{ik}-S_{jk})^2} & i\neq j\end{cases} \tag{5}$$

where, m is the number of attributes; S_{ik} is the attribute value of the i-th row and k-th column [19]. By formula (5), the matrix R can be obtained as Table 3.

Table 2. The initial data on influencing factors of student scores

Serial number	Education of parents	Attention from parents	Preview before class?	Weekly number of self-studies	Learning interest	Mean time of exchange with tutor	...
1	0.1532	0.3323	0.723	0.5323	0.2321	0.4231	
2	0.324	0.2256	0.2542	0.5142	0.3925	0.2914	
3	0.1865	0.1849	0.258	0.2563	0.263	0.2826	
4	0.074	0.3135	0.7432	0.2123	0.3789	0.2835	
5	0.332	0.2536	0.269	0.5112	0.3896	0.4231	
6	0.1562	0.3023	0.7253	0.523	0.3965	0.2915	
7	0.1365	0.1875	0.7325	0.211	0.3825	0.2865	
9	0.2238	0.2623	0.7229	0.523	0.3932	0.2915	
10	0.2531	0.2836	0.2725	0.5324	0.3825	0.4235	
11	0.312	0.2236	0.2356	0.528	0.3623	0.498	
12	0.152	0.3325	0.7132	0.2176	0.3942	0.2786	
13	0.311	0.261	0.752	0.5239	0.3724	0.4235	
14	0.248	0.245	0.761	0.5256	0.2135	0.425	
...	...	...	...	...	...	...	...

2.3 Clustering

(1) Taking λ=0.817, the influencing factors of student scores were divided into six groups: $\{A_1, A_2\}$; $\{A_3\}$; $\{A_{12}, A_4\}$; $\{A_5, A_6, A_8, A_9, A_7\}$; $\{A_{10}, A_{11}\}$; $\{A_{13}\}$.

Group 1 includes attention from parents, and education of parents;

Group 2 includes NCEE results;

Group 3 includes mean time of exchange with tutor, and monthly number of exchanges with tutor;

Group 4 includes frequency of independent completion of homework, learning interest, weekly nonattendance, preview before class? and weekly number of self-studies [20];

Group 5 includes learning satisfaction, and influence of roommates;

Group 6 includes participation in clubs and student union?

(2) Taking λ=0.773, the influencing factors of student scores were divided into four groups [21]:

$\{A_1, A_2\}$; $\{A_3\}$; $\{A_{10}, A_{12}, A_5, A_4, A_5, A_6, A_8, A_9, A_7\}$; $\{A_{13}, A_{11}\}$.

Group 1 includes attention from parents, and education of parents;

Group 2 includes NCEE results;

Group 3 includes mean time of exchange with tutor, monthly number of exchanges with tutor, frequency of independent completion of homework, learning interest, weekly nonattendance, preview before class? weekly number of self-studies, and learning satisfaction;

Group 4 includes participation in clubs and student union? and influence of roommates.

To sum up, it is important to divide the factors affecting student scores under certain conditions. The greater the λ values, the more refined the divisions, and the better the pertinence. The FCA can excellently divide the influencing factors, facilitating the subsequent screening of the key factors and formulation of countermeasures.

Table 3. The fuzzy similarity matrix

1.00																			
0.17	1.00																		
0.15	0.19	1.00																	
0.18	0.15	0.18	1.00																
0.16	0.18	0.06	0.18	1.00															
0.18	0.16	0.18	0.14	0.16	1.00														
0.16	0.17	0.17	0.07	0.17	0.14	1.00													
0.17	0.15	0.16	0.24	0.18	0.21	0.22	1.00												
0.15	0.14	0.13	0.23	0.14	0.17	0.23	0.32	1.00											
0.19	0.09	0.18	0.17	0.17	0.12	0.18	0.22	0.25	1.00										
0.21	0.23	0.16	0.16	0.15	0.19	0.17	0.16	0.24	0.18	1.00									
0.08	0.22	0.14	0.18	0.14	0.18	0.18	0.19	0.32	0.16	0.14	1.00								
0.17	0.16	0.12	0.17	0.12	0.15	0.21	0.12	0.04	0.15	0.14	0.15	1.00							
0.15	0.15	0.13	0.13	0.18	0.22	0.15	0.14	0.15	0.17	0.08	0.12	0.14	1.00						
0.16	0.18	0.17	0.14	0.21	0.19	0.13	0.18	0.24	0.16	0.21	0.14	0.26	0.16	1.00					
0.18	0.06	0.19	0.12	0.16	0.11	0.17	0.17	0.16	0.12	0.22	0.15	0.17	0.21	0.16	1.00				
0.14	0.13	0.07	0.21	0.08	0.12	0.18	0.16	0.14	0.15	0.19	0.16	0.13	0.17	0.18	0.13	1.00			
0.21	0.18	0.18	0.13	0.13	0.13	0.08	0.22	0.21	0.14	0.17	0.17	0.25	0.18	0.17	0.14	0.16	1.00		
0.17	0.17	0.06	0.14	0.07	0.16	0.17	0.15	0.15	0.22	0.14	0.15	0.13	0.19	0.24	0.16	0.08	0.18	1.00	
0.19	0.16	0.19	0.06	0.18	0.17	0.12	0.17	0.16	0.13	0.13	0.17	0.18	0.14	0.19	0.17	0.19	0.14	0.21	1.00

3. PCA AND CLUSTER ANALYSIS ON IMPACT MECHANISMS

The above section mainly explores the influence of various dimensions (e.g. school, student, teacher, and society) on student scores. Despite the wide scope, the main factors were not identified. Focusing on the correlations between many factors [22], the factor analysis expresses the main information of the original factors with a few extracted factors, making the data more condense. The basic task of factor analysis is to determine factor loading. The following is the factor analysis on the factors affecting student scores, and the interpretation of the analysis results.

3.1 Objects

After interview, observations, and literature review, the authors designed a questionnaire on 13 factors (A_1-A_{13}) in four dimensions (school, student, family, and society) that potentially affect student scores. In the questionnaire, 2-4 choices are provided for each factor [23]. Through cluster sampling, 300 questionnaires were randomly distributed among students of different grades from different colleges. A total of 205 valid questionnaires were returned.

3.2 Survey process and analysis methods

The questionnaire survey was conducted online. Each respondent filled out the questionnaire with the help of his/her classmates. The questionnaires were collected immediately. The survey data were summarized and converted by Excel into a quantitative statistical table, and analyzed on SPSS21.0 and SAS8.0.

3.3 Questionnaire quantification

Through questionnaire quantification, the positive factors were differentiated from negative factors. For each positive factor, the value of each choice increases with the degree of positive impact; for each negative factor, the value of each choice increases with the degree of negative impact. After value assignment, a quantitative statistical table (Table 4) was obtained, which contains 13 factors and 205 samples.

Table 4. The results of factor clustering

	ANOVA					
	Clustering		Error		F ratio	Sig.
	Sum of squares	Degrees of freedom (DOFs)	Sum of squares	DOFs		
Education of parents	.001	1	.003	235	.151	.735
Attention from parents	.000	1	.001	235	.096	.748
Learning interest	3.487	1	.019	235	158.312	.002
Monthly number of exchanges with tutor	.691	1	.021	235	37.827	.003
Weekly number of self-studies	.016	1	.004	235	2.612	.123
Mean time of exchange with tutor	.002	1	.003	235	.528	.479
Weekly nonattendance	.000	1	.000	235	.061	.825
Learning satisfaction	.002	1	.004	235	.289	.576
Frequency of independent completion of homework	.000	1	.008	235	.058	.851
Preview before class?	.000	1	.002	235	.038	.863
Participation in clubs and student union?	.000	1	.021	235	.002	.948
Influence of roommates	1.049	1	.013	235	36.495	.000
NCEE results	.001	1	.011	235	.078	.729

Since the clusters were selected to maximize the intra-class difference [25], F-test should only be used for descriptive purposes. The measured significance, which was not modified, cannot verify the hypothesis of equal cluster means [26]. The final cluster heads and classes of influencing factors are shown in Tables 5 and 6, respectively.

Table 5. The final cluster heads

Factors	Classes	
	1	2
Education of parents	.1924	.1978
Attention from parents	.2428	.2536
Learning interest	.7235	.4612
Monthly number of exchanges with tutor	.3352	.4489
Weekly number of self-studies	.3467	.3336
Mean time of exchange with tutor	.3498	.3461
Weekly nonattendance	.2625	.2694
Learning satisfaction	.3607	.3513
Frequency of independent completion of homework	.3628	.3617
Preview before class?	.2498	.2486
Participation in clubs and student union?	.5356	.5348
Influence of roommates	.4756	.3336
NCEE results	.4215	.4224

Table 6. The classes of influencing factors

Classes	Factors
1	A1, A2
2	A3
3	A4, A12
4	A5, A6, A7, A8, A9
5	A10, A11
6	A13

3.4 Extraction of influencing factors

Cluster analysis was combined with PCA for dimensionality reduction. The combined strategy is suitable for dimensionality reduction problems with multiple factors, because it can measure the significance of principal components, and eliminate the multicollinearity between variables.

Here, the 13 presupposed factors are subject to clustering in R. Then, each class of factors was subject to PCA to extract the principal components. Finally, the principal components of each class were merged into the main factors affecting college student scores [24].

3.5 Clustering results

Ward's hierarchical cluster analysis was performed on all factors. According to the number of classes suggested by multiple statistics, the final number of classes was determined as 4. The results of factor clustering are displayed in Table 4.

The linear relationship between each factor and the presupposed factors can be obtained as:

$$X1 = 0.778A3 + 0.761A4 + 0.611A11 + 0.429A12$$
$$......$$
$$X2 = 0.18A9 - 0.464A12 + 0.835A5 + 0.29A6 - 0.265A3 - 0.168A9 \tag{6}$$

3.6 PCA and factor analysis

3.6.1 Principle of principal component extraction

The principal components of each class were extracted by the principle that the eigenvalue of vector correlation coefficient matrix must be greater than 1.

3.6.2 PCA results

The PCA results are displayed in Tables 7 and 8. As shown in Table 8, the first 6 principal components, whose eigenvalues are greater than 1, cumulatively contribute to 71.123% of the variance. With large eigenvalues, the first 6 principal components explain 71.123% of the variance. Therefore, these principal components were selected to evaluate the factors that affect the linear algebra scores of college students.

Table 7. The name of factors and common factor variance

Code	Factor	Initial	Extracted
A1	Attention from parents	1.000	.763
A2	Education of parents	1.000	.812
A3	NCEE results	1.000	.656
A4	Mean time of exchange with tutor	1.000	.536
A5	Frequency of independent completion of homework	1.000	.637
A6	Learning interest	1.000	.482
A7	Weekly nonattendance	1.000	.528
A8	Preview before class?	1.000	.673
A9	Weekly number of self-studies	1.000	.189
A10	Learning satisfaction	1.000	.637
A11	Influence of roommates	1.000	.581
A12	Monthly number of exchanges with tutor	1.000	.624
A13	Participation in clubs and student union?	1.000	.613
	Extraction method: PCA		

Table 8. The total variance explained

Component	Initial eigenvalues			Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	1.699	13.132	14.561	1.765	13.487	13.512	1.812	13.312	13.265
2	1.324	11.035	32.425	1.323	9.951	23.456	1.225	9.235	22.628
3	1.275	8.512	45.953	1.258	9.492	32.846	1.189	9.218	31.872
4	1.189	9.217	61.041	1.212	9.213	42.236	1.171	8.891	40.769
5	1.112	8.354	69.325	1.081	8.312	50.256	1.172	8.982	49.793
6	1.078	8.297	70.026	1.074	8.225	58.635	1.145	8.236	58.498
7	.974	7.521	79.314						
8	.935	7.342	81.295						
9	.883	6.698	84.265						
10	.821	6.362	89.334						
11	.768	5.569	92.997						
12	.656	5.891	98.035						
13	1.698	13.256	12.995	1.864	12.995	13.975	1.641	13.235	13.672
	Extraction method: PCA								

3.6.3 Results of factor analysis

By formula (1), the main factors that constitute each influencing factor were extracted in descending order of absolute value of factor coefficients, and named after the rank of that value (Table 9). By varimax with Kaiser normalization, the new factor loadings of the 13 influencing factors were obtained on the six factors. As shown in Table 9, factor 1 is dominated by A1, and A2; factor 2 is dominated by A3; factor 3 is dominated by A4, and A12; factor 4 is dominated by A6, A5, A7, A8, and A9; factor 5 is dominated by A10, and A11; factor 6 is dominated by A13. Factor 1 mainly reflects the conditions of parents, factor 2 mainly reflects NCEE results, factor 3 mainly reflects the exchange between student and tutor, factor 4 mainly reflects the homework completion and self-learning, factor 5 mainly reflects the roommate influence and learning satisfaction, and factor 6 mainly reflects the environment at school. Therefore, the six factors were referred to as family factor, examine factor, exchange factor, learner factor, classmate factor, and campus factor (as shown in Table 10).

Through exploratory factor analysis, six potential factors were found from the 13 presupposed influencing factors: family factor, exam factor, exchange factor, learner factor, classmate factor, and campus factor. There is no cross-influence between them, that is, each influencing factor is only affected by one potential factor. Hence, the six potential factors are the main factors affecting student scores.

Table 9. The rotated component matrix

		Rotated component matrix[a]
Code	Factors	Components
		1	2	3	4	5	6
A1	Attention from parents	.886	-.054	.026	-.029	-.038	.112
A2	Education of parents	.879	-.022	.085	.055	-.039	.056
A3	NCEE results	.112	.731	.043	.019	-.131	-.231
A4	Mean time of exchange with tutor	-.128	.232	.665	.051	.153	.225
A5	Frequency of independent completion of homework	.245	-.259	.112	.556	-.049	-.218
A6	Learning interest	-.018	-.078	.743	.658	.036	.059
A7	Weekly nonattendance	.016	-.013	-.645	.774	.058	-.051
A8	Preview before class?	-.161	.004	-.256	.679	-.119	.258
A9	Weekly number of self-studies	.033	.149	.239	.609	-.048	-.386
A10	Learning satisfaction	-.025	.278	.013	-.159	.731	.069
A11	Influence of roommates	.167	-.221	-.064	.358	.535	.076
A12	Monthly number of exchanges with tutor	.122	.288	.825	.084	-.523	.369
A13	Participation in clubs and student union?	.033	.016	.142	.051	-.008	.769
Extraction method: PCA; Rotation method: Varimax with Kaiser normalization; a. Rotation converged in 9 iterations.

Table 10. The main influencing factors and factor names

Code	Influencing factors	Factor names
X1	A1, A2	Family factor
X2	A3	Exam factor
X3	A4, A12	Exchange factor
X4	A5, A6, A7, A8, A9	Learner factor
X5	A10, A11	Classmate factor
X6	A13	Campus factor

4. ANOVA ON IMPACT MECHANISMS

This section performs ANOVA on the elements of each potential factor, aiming to find the element that contributes the greatest to the potential factor. The ANOVA results provide valuable reference for improving teaching quality and student scores.

4.1 Multi-way ANOVA of family factor

As shown in Table 11 below, A1 had the most significant effect in family factor, i.e. A1 has greater impact on student scores than A2.

4.2 Multi-way ANOVA of exchange factor

As shown in Table 12 below, A12 had the most significant effect in exchange factor, i.e. A12 has greater impact on student scores than A4.

4.3 Multi-way ANOVA of learner factor

As shown in Table 13 below, A15 had the most significant effect in learner factor, i.e. A5 has greater impact on student scores than A5, A6, A7, A8, and A9.

4.4 Multi-way ANOVA of classmate factor

As shown in Table 14 below, A10 had the most significant effect in classmate factor. But there is no reason to conclude that A11 does not have a significant effect on student scores.

4.5 One-way ANOVA of campus factor

As shown in Table 15 below, there is no evidence that A13 has or does not have significant effect on student scores.

4.6 One-way ANOVA of exam factor

As shown in Table 16 below, there was significant difference in intra-subject means. Under the significance level of 0.05, the F ratio was 0.436, greater than the corresponding p-value of 0.009. Hence, the original hypothesis that NCEE results have a significant effect on student scores was rejected.

4.7 Discussion

Through the above ANOVAs, it is learned that, among the presupposed factors, A1, A12, A5, A10, and A3 are the leading factors affecting college student scores. Further ANOVA reveals that A3 and A10 are the two top influencing

factors of college student scores. The above analysis shows that the improvement of teaching quality requires the concerted efforts from the school, teachers, and students. To promote the development of the school and students, the school management should invest more in hardware facilities and soft power, creating a favorable learning, working, and living environment for students and teachers. The teachers should teach students in accordance with their aptitude, continuously improve their teaching skills, and adopt various teaching methods and means, making the students more interested, and proactive in learning. The students must concentrate their energy in learning, and lay a solid foundation for advanced professional courses.

Table 11. The intra-cluster effects (intra-subject effect test)

Intra-subject effect test					
Dependent variable: V15					
Source	Type III sum of squares	DOFs	Mean square	F ratio	Sig.
Corrected model	.018[a]	5	.004	.786	.598
Intercept	19.281	1	18.734	3368.562	.000
Education of parents	38.567	2	14.102	.365	.534
Attention from parents	46.335	2	36.384	0.525	.0.875
Education of parents * Attention from parents	.000	0	.	.	.
Error	1.050	212	.005		
Total	249.683	235			
Corrected total	100.152	234			
a. R^2 = .019(adjusted R^2 = -.006)					

Table 12. The intra-cluster effects (intra-subject effect test)

Intra-subject effect test					
Dependent variable: V15					
Source	Type III sum of squares	DOFs	Mean square	F ratio	Sig.
Corrected model	.018[a]	9	.002	.412	.878
Intercept	16.632	1	16.668	2986.334	.000
Monthly number of exchanges with tutor	.004	2	.002	.386	.759
Mean time of exchange with tutor	.003	2	.002	.787	.758
Monthly number of exchanges with tutor * Mean time of exchange with tutor	.011	4	.003	.534	.765
Error	1.048	199	.005		
Total	18.827	235			
Corrected total	1.134	234			
a. R^2 = .028 (adjusted R^2 = -.033)					

Table 13. The intra-cluster effects (intra-subject effect test)

Intra-subject effect test					
Dependent variable: V15					
Source	Type III sum of squares	DOFs	Mean square	F ratio	Sig.
Corrected model	.324[a]	83	.004	.586	.897
Intercept	9.450	1	9.380	1531.528	.000
Weekly number of self-studies	.005	2	.003	.456	.636
Frequency of independent completion of homework	.006	2	.003	.723	.627
Weekly nonattendance	.001	2	.001	.096	.935
Preview before class?	.006	2	.003	.512	.669
Learning interest	.000	0	.	.	.
Error	.773	189	.006		
Total	18.827	235			
Corrected total	1.134	232			
a. R^2 = .324(adjusted R^2 = -.190)					

Table 14. The intra-cluster effects (intra-subject effect test)

Intra-subject effect test					
Dependent variable: V15					
Source	Type III sum of squares	DOFs	Mean square	F ratio	Sig.
Corrected model	.41[a]	8	.005	.887	.534
Intercept	15.227	1	15.323	2823.719	.000
Influence of roommates	.016	2	.007	1.523	.261
Learning satisfaction	.013	2	.005	1.891	.345
Influence of roommates * Learning satisfaction	.012	4	.003	.628	.721
Error	1.031	196	.005		
Total	18.827	235			
Corrected total	1.134	232			
a. R^2 = .041 (adjusted R^2 = -.003)					

Table 15. The intra-cluster effects (intra-subject effect test)

Source	Type III sum of squares	DOFs	Mean square	F ratio	Sig.
Intra-subject effect test					
Dependent variable: V15					
Corrected model	.008ᵃ	1	.008	1.468	.236
Intercept	17.835	1	16.256	3356.724	.000
Participation in clubs and student union?	.009	1	.008	0.537	.209
Error	1.231	199	.005		
Total	18.827	235			
Corrected total	1.134	232			
a. R^2 = .006(adjusted R^2 = .002)					

Table 16. The intra-cluster effects (intra-subject effect test)

Source	Type III sum of squares	DOFs	Mean square	F ratio	Sig.
Intra-subject effect test					
Dependent variable: V15					
Corrected model	.010ᵃ	2	.005	.436	.009
Intercept	14.489	1	13.624	2598.216	.000
NCEE results	.011	2	.005	1.923	.009
Error	1.137	198	.005		
Total	18.827	235			
Corrected total	1.134	232			
a. R^2 = .008(adjusted R^2 = -.002)					

5. CONCLUSIONS

Drawing on the relevant literature, this paper attempts to clarify the factors that truly affect the college student scores. A total of 13 potential factors were selected, including learning interest, frequency of independent completion of homework, and dormitory atmosphere. Then, a relevant questionnaire survey was conducted among students of different grades from different colleges. FCA on the collected data did not discover good correlations between these factors. Then, the PCA was performed on the survey data, revealing good correlations between the factors. That is, the 13 potential factors could be divided into six groups, and 7 factors will change with the six groups. Finally, the six groups were treated through ANOVA. The results show that the six groups do not interfere with each other. Hence, the fix groups are the main influencing factors of college student scores: family factor, Exam factor, exchange factor, learner factor, classmate factor, and campus factor.

Of course, there are several limitations of this research: the impact mechanisms of the six groups of factors were not clarified; the subjects of the questionnaire survey are too small compared with the total number of college students (37 million) in China. To make up for the limitations, the future research needs to collect a massive number of representative samples, and measure the exact impact of each factor that affects college student scores.

REFERENCES

[1] Gbollie, C., Keamu, H.P. (2017). Student academic performance: The role of motivation, strategies, and perceived factors hindering Liberian junior and senior high school students learning. Education Research International, 2017. https://doi.org/10.1155/2017/1789084.

[2] Cheng, W., Ickes, W., Verhofstadt, L. (2012). How is family support related to students' GPA scores? A longitudinal study. High Educ, 64: 399-420. https://doi.org/10.1007/s10734-011-9501-4

[3] Wang, D.F. (1992). Influence of psychological control source tendency on blame and justification: Further evidence. Journal of Psychology, 1992(2): 174-181.

[4] Hazrati-Viari, A., Rad, A.T., Torabi, S.S. (2012). The effect of personality traits on academic performance: The mediating role of academic motivation. Procedia -Social and Behavioral Sciences, 32: 367-371. https://doi.org/10.1016/j.sbspro.2012.01.055

[5] Nechita, F., Alexandru, D.O., Turcu-Ştiolică, R., Nechita, D. (2015). The influence of personality factors and stress on academic performance. Current Health Sciences Journal, 41(1): 47-61. https://doi.org/10.12865/CHSJ.41.01.07

[6] Yigermal, Y.M. (2017). Determinant of academic performance of under graduate students: In the cause of Arba Minch University Chamo Campus. Journal of Education and Practice, 8(10): 155-166.

[7] Musah, M.B., Ali, H.B.M., Al-Hudawi, S.H.V., Tahir, L.M., Daud, K.B., Hamdan, A.R. (2015). Determinants of students' outcome: A full-fledged structural equation modelling approach. Asia Pacific Educ, 16: 579-589. https://doi.org/10.1007/s12564-015-9396-3

[8] Olango, M. (2016). Mathematics anxiety factors as predictors of mathematics self-efficacy and achievement among freshmen science and engineering students. African Educational Research Journal, 4(3): 109-123.

[9] Zhu, Z.B., Chen, L.L., Jin, Z.G. (2017). Analysis of influencing factors of classroom silence of college students: From the perspective of implicit theory. Science of University Education, 6: 50-56. https://doi.org/10.3969/j.issn.1672-0717.2017.06.010

[10] Xiao, Q.H., Zhang, L.R., Shi, E.H. (2015). Statistical analysis of influencing factors of mathematics achievement of college students in different grades. Journal of Mathematics Education, 24(4): 53-56.

[11] Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L.D.F., Rodrigues, F.A.

(2019). Clustering algorithms: A comparative approach. Research Article, 15. https://doi.org/10.1371/journal.pone.0210236

[12] Zabihi, S.M., Akbarzadeh-T, M.R. (2012). Generalized fuzzy C-means clustering with improved fuzzy partitions and shadowed sets. Research Article Open Access, 2012. https://doi.org/10.5402/2012/929085

[13] Scitovski, R., Vidović, I., Bajer, D. (2016). A new fast fuzzy partitioning algorithm. Expert Systems with Applications, 51: 143-150. https://doi.org/10.1016/j.eswa.2015.12.034

[14] Espitia, H., Soriano, J., Machón, I., López, H. (2019). Design Methodology for the implementation of fuzzy inference systems based on Boolean relations. Electronics, 8(11): 1243. https://doi.org/10.3390/electronics8111243

[15] Peng, Y., Ding, S.L. (2009). Cluster analysis technology based on attribute reduction. Computer Engineering and Applications, 45(9): 138-140+195.

[16] Jain, A., Sheel, S., Bansal, K. (2016). Constructing fuzzy membership function subjected to GA based constrained optimization of fuzzy entropy function. Indian Journal of Science and Technology, 9(43): 1-10. https://doi.org/10.17485/ijst/2016/v9i43/104401.

[17] Fu, Y., Pan, S.Y. (2013). Application of fuzzy clustering in customer relationship management. Software guide, 12(10): 49-51.

[18] Peng, Y., Nie, C.Q., Yu, S.L. (2006). Association rules mining strategy based on reduced data sets. Computer Engineering and Applications, 11: 169-172.

[19] Li, Y., Xia, D., Dan, Z. (2013). Performance evaluation index system of knowledge management operation in high-tech industrialization. Statistics and Decision-Making, 4: 21-24.

[20] Ciaramella, A., Nardone, D. Staiano, A. (2020). Data integration by fuzzy similarity-based hierarchical clustering. BMC Bioinformatics, 21: 350. https://doi.org/10.1186/s12859-020-03567-6

[21] Noë, R., Sluijter, A.A. (1995). Which adult male savanna baboons form coalitions. International Journal of Primatology, 16(2): 77-105. https://doi.org/10.1007/BF02700154

[22] Yang, Y.L. (2008). Research on factors influencing college students' life satisfaction. Journal of Science and Technology Innovation, 16: 189-200.

[23] Zhang, Y.H., Shang, Y.M., Ji, H.T. (2018). Attribution and reflection on achievement of undergraduates with difficulty in probability theory and mathematical statistics. Journal of Capital Normal University (Natural Science Edition), 39(1): 8-12.

[24] Jolliffe, I.T., Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions Mathematical Physical & Engineering Sciences, 374(2065): 20150202. https://doi.org/10.1098/rsta.2015.0202.

[25] González, B., López, A., García, R. (2008). Supreme Audit Institutions and their communication strategies. International Review of Administrative Sciences. https://doi.org/10.1177/0020852308095312

[26] Gómez-Adorno, H., Martín-del-Campo-Rodríguez, C., Sidorov, G., Alemán, Y., Vilariño, D., Pinto, D. (2018). Hierarchical clustering analysis: The best-performing approach at PAN 2017 author clustering task. In: Bellot P. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2018. Lecture Notes in Computer Science, 11018. Springer, Cham. https://doi.org/10.1007/978-3-319-98932-7_20