Education Insights

Article

Analysis of Test Scores Considering the Influence of Question Difficulty: A Case from Bridge Seismic and Wind Resistance Course

Yafeng Li [1], Changbai Wang [1,*], Qiang Wang [1], Yuan Song [1], Yuxuan Wang [1], Gonghong Hu [1], Pu Yuan [1] and Wei He [1]

[1] School of Civil Engineering and Architecture, Anhui University of Science and Technology, Huainan, Anhui, 232001, China

* Correspondence: Changbai Wang, School of Civil Engineering and Architecture, Anhui University of Science and Technology, Huainan, Anhui, 232001, China

Abstract: Bridge Seismic and Wind Resistance Design is a newly offered elective course for the civil engineering major. Due to a lack of experience in test design for this new course and the scattered nature of its knowledge points, the difficulty of some test questions in the final examination deviated, which affected the accurate subsequent assessment of teaching effect subsequently. Therefore, this paper first analyzes the influence of test question difficulty on its scores. Then, the Set Pair Analysis (SPA) theory was introduced, and the score rate was selected as the indicator to measure the contribution of each question with different difficulty degrees to the overall test scores evaluation. A comprehensive test scores analysis model that can consider the test question difficulty is established. Finally, the validity of this test scores analysis model was explored through the class performance of the past year, which can provide theoretical guidance and data support for continuous teaching reform and quality assessment.

Keywords: test scores; question difficulty; set pair relation criterion; comprehensive analysis model; newly offered course

Published: 17 January 2026

1. Introduction

Under the policies and advocacy of the Ministry of Education, many universities have begun to focus on social development needs to reform professional curricula, adding a series of characteristic elective courses combined with the university's teaching and research characteristic platforms to cultivate specialized talents with distinct characteristics [1-2]. The authors' university has added a large number of major elective courses in the field of bridge and road construction for civil engineering major since last year, including *Bridge Seismic and Wind Resistance Design*. After two years of undergraduate teaching, the authors found that this course is highly theoretical, practical, and closely related to current norms, regulations and standards, when using test papers for assessment, factors such as insufficient experience in question setting, inappropriate selection of assessment points, and uneven question difficulty led to extremely high or low scores on some questions. The difficulty level of test questions directly affects the overall score level and distribution of candidates, also making the paper score unable to truly reflect the students' overall mastery of the course knowledge [3-4].

The reanalysis and evaluation of test scores essentially involves assessing the effectiveness of each question in measuring students' overall mastery of the course knowledge. This problem is influenced by both the students' subjective knowledge mastery and the objective difficulty of the questions, making it a complex uncertain system problem [5]. Set Pair Analysis (SPA) theory is an analytical method for uncertain system problems, widely used in fields such as data mining, condition assessment, and pattern recognition [6]. Many scholars have used SPA theory to analyze exam scores, including course grades and their developmental trends [7]. These research efforts demonstrate the effectiveness of SPA theory in solving the uncertain system problem of analyzing university student grades. However, the above studies used the traditional method of summing scores for each question when obtaining the first-hand data needed for analysis, ignoring the impact of varying question difficulty on students' paper scores. An excessively high proportion of low-difficulty, high-point questions can inflate paper scores compared to students' actual mastery of course knowledge, and vice versa [8]. Therefore, it is necessary to measure the differential contribution to comprehensive grades due to differences in question difficulty to obtain more comprehensive and accurate overall grades.

Taking the newly added elective course *Bridge Seismic and Wind Resistance Design* as an example, considering that insufficient teaching and assessment experience led to some questions being too difficult or too easy, after the preliminary analysis of the scores of student performance, using the score rate as an indicator to reanalyze the information contained in the paper scores. Based on the distribution characteristics of the score rate, a set pair relationship criterion is established to describe the differential contribution of each question to the comprehensive evaluation. Ultimately, a comprehensive performance analysis model that can consider the influence of test question difficulty is obtained.

2. Question Characteristics and Score Distribution

Bridge Seismic and Wind Resistance Design as a newly added major elective course for civil engineering, was assessed using an open-book examination. The question types, point values, and assessment difficulty are detailed in Table 1. Overall, short-answer questions had the lowest difficulty, students could directly look up relevant content from the textbook to answer them, and this question type had the highest point value. Single-choice and multiple-choice questions could basically be found in the textbook, but the knowledge points were scattered, and it is impossible to look up every question within the limited exam time, resulting in moderate difficulty overall. Material-based questions included one material analysis question and one calculation question, requiring students not only to master relevant calculation theories proficiently but also to apply the course knowledge in combination with actual engineering projects, thus having the highest difficulty.

Table 1. Question types and answering difficulty.

Question No.	Type/Points	Source and difficulty
1	Single-choice/10	1) Directly from the textbook; 2) Moderate difficulty.
2	Multiple-choice/20	1) Basically directly from the textbook; 2) Moderate difficulty, but missing or wrong selections lower the score rate.
3	Short-answer/42	1) Can be directly excerpted and summarized from the textbook; 2) Low question difficulty.
4	Material-based/28	1) Material analysis question cannot be answered directly from the book, requires students to combine classroom knowledge. Higher difficulty. 2) Calculation question has similar exercises in the textbook. Lower difficulty given sufficient proficiency.

After the exam, 72 valid answer sheets were collected. Since each question had different point values, to facilitate subsequent performance analysis, the score for each question was divided by its total points to obtain the score rate for each question. Its distribution is shown in Figure 1. A normal distribution function was used to fit the statistical results, as shown in Figure 1 and Table 2.

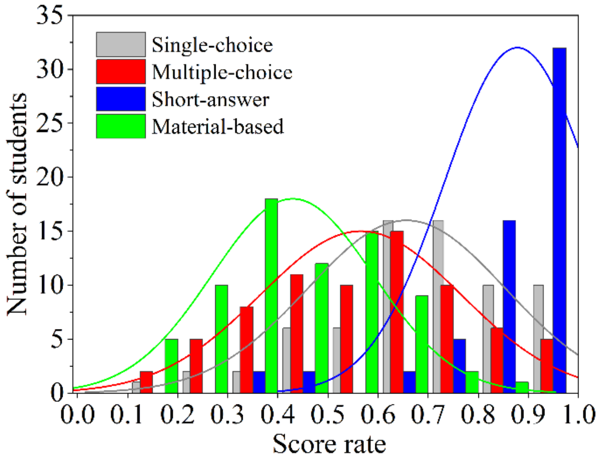

Figure 1. Score rate distribution for various types of questions.

Table 2. Normal distribution model parameters for each question's scores.

Question type	Mean(μ)	Standard deviation(σ)
Single-choice	0.657	0.197
Multiple-choice	0.566	0.201
Short-answer	0.878	0.149
Material-based	0.430	0.161

In terms of the mean score rate, Short-answer > Single-choice > Multiple-choice > Material-based. This is because short-answer questions were mostly key teaching content, easy to find sources for in the book and answer, thus having the highest score rate. Single-choice questions followed, slightly higher than multiple-choice, as these types were also relatively easy to find knowledge points for in the book, resulting in acceptable score rates. However, multiple-choice questions were prone to score loss due to missing or wrong selections, making their score rate lower than single-choice. Material-based questions had the lowest score rate, firstly because the material analysis question required students to discuss combining actual situations, an area where student ability is generally lacking; secondly, calculations required a certain proficiency. Attempting to understand and calculate for the first time during the exam easily reduced the score rate due to time constraints.

In terms of the dispersion of scores (standard deviation), Multiple-choice > Single-choice > Material-based > Short-answer. For single and multiple-choice questions where answers could not be confirmed from the book, individual differences in knowledge mastery led to greater score dispersion. Comparatively, multiple-choice questions better reflected differences in knowledge mastery; wrong or missing selections caused greater fluctuations in scores for these questions. Therefore, the score dispersion for multiple-choice was greater than for single-choice. Material-based questions tested students' ability to apply theoretical knowledge to practical problems and their proficiency in theoretical calculation methods. Most students still lacked ability in these two aspects, leading to generally low score rates with little fluctuation. Short-answer questions could basically be answered by excerpting and integrating from the textbook, which most students could do effectively, resulting in high score rates and low dispersion.

3. Comprehensive Performance Analysis Model Considering Test Question Difficulty

3.1. Set Pair Analysis Theory

Set Pair Analysis (SPA) theory is a method for analyzing uncertain system problems. This theory constructs a set pair from two related objects A and B, analyzing the certainty and uncertainty relationships between the set pair through *N* characteristics. The *N* characteristics are divided into identity, discrepancy, and contrary relationships, obtaining the connection number $\mu_{A\text{-}B}$ for the set pair *H*(A-B), as shown in Equation (1).

$$\mu_{A\text{-}B}=\frac{S}{N}+\frac{F}{N}i+\frac{P}{N}j \tag{1}$$

In the equation, *S*, *F* and *P* are the number of characteristics among the *N* characteristics where the set pair relationship is identity, discrepancy and contrary, respectively, with $S+F+P=N$; *i* is the discrepancy degree coefficient, taking values in the interval (-1, 1); *j* is the contrary degree coefficient, often taken as -1.

Generally, the greater the number of characteristics in the identity relationship, the larger the connection number between the two objects constituting the set pair, and the more the set pair shows a relatively certain positive correlation. The greater the number of characteristics in the contrary relationship, the smaller the connection number, and the more the set pair shows a relatively certain negative correlation. The greater the number of characteristics in the discrepancy relationship, depending on the value of the discrepancy degree coefficient *i*, the connection number shows either positive or negative changes, and the set pair shows a relatively uncertain correlation.

3.2. Set Pair Relationship Criterion

From the basic introduction of SPA theory above, it is clear that determining the set pair relationship criterion, that is, judging whether a specific set pair characteristic belongs to identity, discrepancy, or contrary, is the core of conducting set pair analysis. The reasonableness of the set pair relationship criterion directly affects the judgment of the set pair relationship to which a characteristic belongs, thereby determining the value of the connection number. The conventional trisection method ignores the objective distribution pattern of set pair characteristics when determining set pair relationships, especially when the distribution of characteristics is uneven. The trisection method may categorize most characteristics into one set pair relationship, ultimately reducing the effectiveness of SPA. Therefore, this paper proposes a new method for dividing set pair relationships based on the distribution characteristics of set pair characteristics in the practical problem of test performance analysis.

Figure 2 plots the relationship curve between the mean score rate and its standard deviation. It can be seen that when the score rate is too high or too low, scores tend to be concentrated, playing a smaller role in distinguishing performance and assessing teaching effectiveness. Such questions have a contrary relationship with performance evaluation, and their weight coefficient in the composite grade evaluation should be reduced. When the score rate is moderate, score dispersion is stronger, playing a greater role in distinguishing performance and assessing teaching effectiveness. Such questions have an identity relationship with performance evaluation, and their weight coefficient in the composite grade evaluation should be larger. When the score rate falls between the above two cases, its role in distinguishing performance and assessing teaching effectiveness has a certain degree of uncertainty, showing a discrepancy relationship with the overall grade evaluation, and its weight in the composite grade evaluation should lie between the former two.

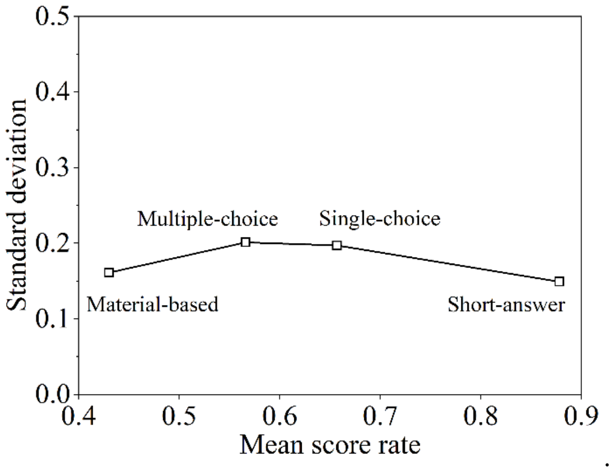

Figure 2. Relationship curve between mean score rate and standard deviation.

Given that the score rate basically satisfies a normal distribution, and combined with the above analysis, a method for determining the set pair relationship criterion based on the score rate is proposed, as shown in Figure 3. When the score rate is within the interval [μ-σ, μ+σ], the question is considered helpful for performance distinction and teaching effectiveness assessment and is classified as having an identity relationship in the overall grade evaluation. When the score rate is within (0, μ-2σ) and (μ+2σ, 100%), the question is considered detrimental to performance distinction and teaching effectiveness assessment and is classified as having a contrary relationship in the overall grade evaluation. When the score rate is within [μ-2σ, μ-σ) and (μ+σ, μ+2σ], the question's role in performance distinction and teaching effectiveness assessment is considered between the above two, and it is classified as having a discrepancy relationship in the overall grade evaluation.

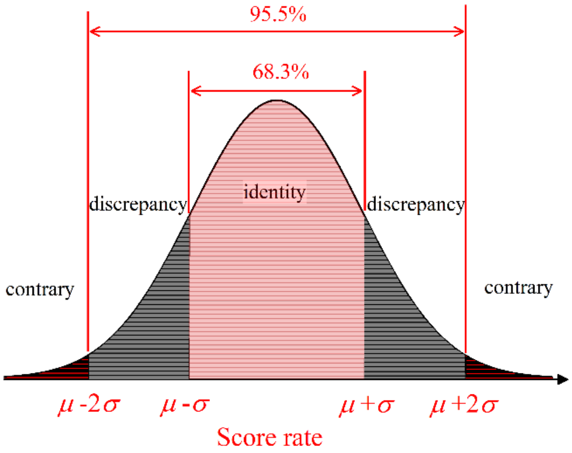

Figure 3. Set pair relationship criterion based on score rate distribution characteristics.

3.3. Comprehensive Performance Evaluation Model

Figure 4 shows the distribution of student scores converted to percentages. Overall, this distribution satisfies a normal distribution pattern. Simultaneously, the overall score distribution is skewed to the right, which is due to the open-book assessment and the high score rates on some questions. The mean score rate for all questions was 0.633, with a standard deviation of 0.241.

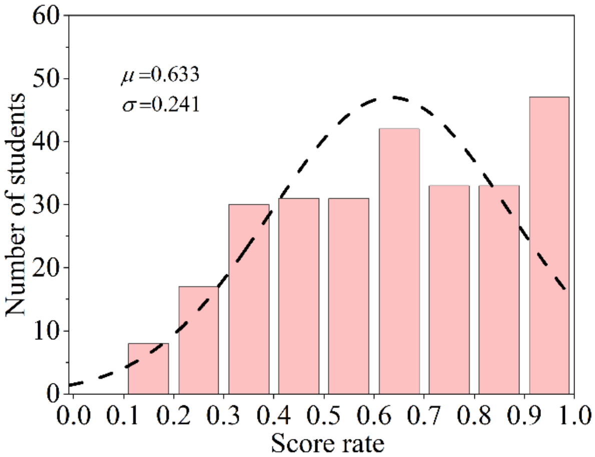

Figure 4. Normal distribution statistical results for score rate.

Based on the set pair relationship division standard in Figure 3, the set pair relationship criterion is obtained as follows:

$$\begin{cases} \&p\in[0.392,0.874]\ \&\&\&\ \text{indentity} \\ \&p\in[0.151,0.392]\cup[0.874,1.115]\ \&\&\&\ \text{discrepancy} \\ \&p\in[-\infty,0.151]\cup[1.151,+\infty]\ \&\&\&\text{contrary} \end{cases} \quad (2)$$

The number of students (out of 72) whose score rate for each question fell into the identity, discrepancy, and contrary categories was counted to obtain the connection degree μ for each question, as shown in Table 3. With the discrepancy degree coefficient *i* taken as 0.5 and the contrary degree coefficient *j* as -1, the connection numbers for single-choice, multiple-choice, short-answer, and material-based questions were calculated as 0.854, 0.813, 0.667, and 0.736, respectively.

Table 3. Set pair connection degree and score adjustment coefficient for each question.

Question Type	Statistics of Set Pair Characteristics per Question			Connection Degree	Weight Coeff. w_k	Score Adj. Coeff. γ_k
	Identity	Discrepancy	Contrary			
Single-choice	54	17	1	$\frac{54}{72}+\frac{17}{72}i+\frac{1}{72}j$	0.278	1.113
Multiple-choice	51	19	2	$\frac{51}{72}+\frac{19}{72}i+\frac{2}{72}j$	0.265	1.059
Short-answer	24	48	0	$\frac{24}{72}+\frac{48}{72}i+\frac{0}{72}j$	0.217	0.869
Material-based	43	26	3	$\frac{43}{72}+\frac{26}{72}i+\frac{3}{72}j$	0.240	0.959

The magnitude of the connection number reflects the weight of that question in the performance analysis. The larger the connection number, the greater the role of that question in distinguishing performance and assessing teaching effectiveness, and the larger its weight coefficient in the comprehensive grade should be. Conversely, the smaller the connection number, the smaller the role of that question in distinguishing performance and assessing teaching effectiveness, and its weight coefficient in the comprehensive grade should be smaller. Therefore, based on the connection numbers, the scores for each question can be adjusted. The following equations are used to normalize the connection numbers to obtain the weight coefficient w_k and the score adjustment coefficient γ_k for each question in the comprehensive evaluation.

$$w_k=\frac{e_k}{\sum_{t=1}^{4}\mu_t} \quad (3)$$

$$\gamma_k=4w_k \quad (4)$$

Where μ_k and μ_t are the connection numbers of the *k*-th question and the *i*-th question, respectively.

The score adjustment coefficients for short-answer and material-based questions are less than 1, because the score distributions for these questions are concentrated, which is unfavorable for performance distinction and overall assessment of teaching effectiveness. The score adjustment coefficients for single-choice and multiple-choice questions are greater than 1, indicating their better role in performance distinction and overall assessment of teaching effectiveness.

Based on the score adjustment coefficient γ_k, the score for each individual question is adjusted, i.e., the original score of the question is multiplied by the adjustment coefficient, and then summed to obtain the comprehensive grade. A comparison between the adjusted comprehensive grade and the conventionally summed grade is shown in Figure 5. The comparison shows that after adjustment, the average student grade decreased from 66.8 points before modification to 62.9 points, effectively mitigating the "inflated score" problem caused by the high proportion and low difficulty of short-answer questions prior to adjustment. This makes the paper score more truly and comprehensively reflect the students' overall mastery of the course knowledge. Furthermore, the original grade distribution, while overall normal, was more skewed to the right, which is a sign of easier test questions. The adjusted grade distribution presents a more balanced normal distribution on both sides, which is a sign of appropriately difficult questions and more consistent with the distribution of student grades under conditions of suitable difficulty.

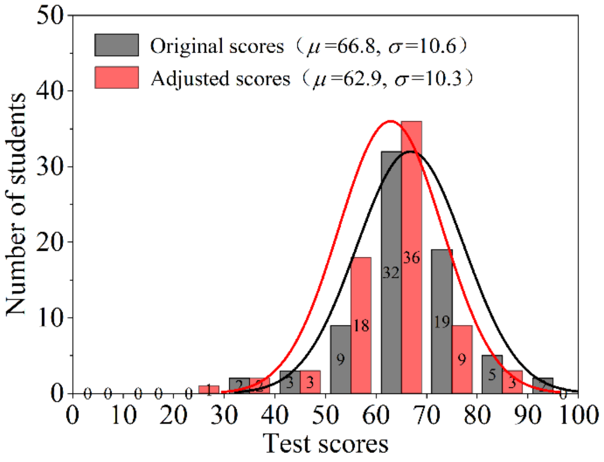

Figure 5. Test scores distribution before and after adjustment.

4. Conclusion

Addressing the situation of insufficient experience in course assessment for the newly established *Bridge Seismic and Wind Resistance Design* course in civil engineering, this paper selected the score rate as an indicator to evaluate test question difficulty, constructed a new set pair relationship criterion to measure the differential contribution of individual questions to the overall grade evaluation due to varying difficulty levels, and established a performance analysis model that can consider test question difficulty. The main conclusions are as follows:

(1) When the score rate for an individual question is too high or too low, scores tend to be concentrated, which is unfavorable for performance distinction and overall assessment of teaching effectiveness. When the score rate is moderate, score discrimination is higher, effectively distinguishing students' learning situations and facilitating the overall assessment of teaching effectiveness.

(2) Based on the normal distribution characteristics of the score rate, the role of an individual question in overall grade evaluation was divided into three types: identity,

discrepancy, and contrary. Based on the connection number, the score adjustment coefficient for individual questions was obtained, clarifying the role of test question difficulty in grade evaluation.

(3) The modified comprehensive grade effectively reduced the role of questions that were too difficult or too easy in the comprehensive grade evaluation, more accurately reflecting the students' learning status.

Funding: This paper was supported by the National-Level First-Class Undergraduate Course (2023240791); Provincial Quality Engineering Projects of Anhui Higher Education Institutions (2021kcszsfkc123, 2022xjzlts007, 2022zygzts026, 2023jyxm0332); Key Projects of Education and Teaching Reform Research at Anhui University of Science and Technology (2023xjjy09; 2024jxjyg1010).

References

1. M. Xie, Y. Sh, X. Liu, and H. Zhang, "Innovation and Practice of Civil Engineering Education under the New Engineering Education Perspective," Frontiers in Educational Research, vol. 7, no. 1, pp. 198-203, 2024.
2. J. Mei, "Construction and Practical Research on the Collaborative Education Model of Promoting All-round Development in the Context of Digital Empowerment," 2025.
3. A. Scarlatos, N. Fernandez, C. Ormerod, S. Lottridge, and A. Lan, "Smart: Simulated students aligned with item response theory for question difficulty prediction," In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, November, 2025, pp. 25082-25105. doi: 10.18653/v1/2025.emnlp-main.1274
4. G. Garman, "A Logistic Approach to Predicting Student Success in Online Database Courses," American Journal of business education, vol. 3, no. 12, pp. 1-6, 2010. doi: 10.19030/ajbe.v3i12.959
5. Y. Xu, S. Hartman, G. Uribe, and R. Mencke, "The effects of peer tutoring on undergraduate students' final examination scores in mathematics," Journal of College Reading and Learning, vol. 32, no. 1, pp. 22-31, 2001. doi: 10.1080/10790195.2001.10850123
6. Z. Chunying, L. Fengchun, and W. Jing, "Set pair mathematical model of teaching quality evaluation system and its application," In 3rd International Conference on Human System Interaction, May, 2010, pp. 531-535. doi: 10.1109/hsi.2010.5514516
7. Y. Liang, H. Wang, and W. C. Hong, "Sustainable development evaluation of innovation and entrepreneurship education of clean energy major in colleges and universities based on SPA-VFS and GRNN optimized by chaos bat algorithm," Sustainability, vol. 13, no. 11, p. 5960, 2021.
8. R. Mourad, and J. H. Hong, "Comparison of developmental student outcomes in college level courses using propensity score matching," Journal of Applied Research in the Community College, vol. 24, no. 1, pp. 59-76, 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). The publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.