SYNERGY PUBLISHING PTE. LTD.

Journal of Educational Theory and Management

https://ojs.s-p.sg/index.php/jetm

Analyzing the Difficulty of China's National College Entrance Examination (NCEE) Mathematics Papers: An AHP Approach with Expert Judgment Matrices and Data-Driven Validation——Evidence from 2024 National Papers A, New Curriculum Standards I & II

Yin Hui* **Yunfei Lu**

School of Mathematical Sciences, Guizhou Normal University, Guiyang, Guizhou, 550025, China

ARTICLE INFO

Article history
Received: 15 April 2025
Accepted: 22 April 2025
Published Online: 30 June 2025

Keywords:
AHP difficulty model
Comparative study
Analysis of variance
NCEE mathematics assessment
Educational measurement

ABSTRACT

This study examines the 2024 National College Entrance Examination (NCEE) Mathematics Papers (National Paper A, New Curriculum Standard I and II) through an integrated methodology combining Analytic Hierarchy Process (AHP) theory, analysis of variance (ANOVA), and Python-based computational analysis. We developed an enhanced difficulty assessment framework featuring an AHP-derived integrated difficulty coefficient model that utilizes expert-constructed judgment matrices for precise factor weighting, overcoming limitations of conventional evaluation approaches. Our Python-powered analytical pipeline enabled comprehensive data processing, visualization of difficulty factor distributions, and rigorous statistical testing. ANOVA results demonstrated no statistically significant differences among the three exam versions across seven core difficulty dimensions: contextual factors, parameter factors, operational proficiency, reasoning ability, knowledge coverage, thinking orientation, and cognitive levels, indicating remarkable stability in overall test difficulty.

1. Introduction

Since the early 21st century, difficulty models have remained a prominent research focus in the field of mathematics education (Yang et al., 2022). In 2002, Chinese scholar Bao Jiansheng, building upon Nohara's related theories, proposed the concept of "Integrated Difficulty in Mathematics Curriculum." He developed an integrated difficulty model for worked examples and exercises across five dimensions: exploration, context, computation, reasoning, and knowledge (Bao, 2002). Based on findings from the Qingpu Experiment, Wang Jianpan and colleagues constructed a framework for mathematical cognitive levels in 2014. They refined the original integrated difficulty model into four levels: operation, concept, comprehension-explanation, and analysis-exploration. This model has been widely applied in comparative studies of textbooks (Wang & Bao, 2014). Then, in 2016, Zhang Yi, Wu Xiaopeng, and others further modified the model and adapted it for research on mathematics questions in China's National College Entrance Examination (NCEE). By 2020, they developed an integrated difficulty model for NCEE mathematics questions using the Analytic Hierarchy Process (AHP) theory (Zhang et al., 2016a; Zhang &

**Corresponding Author:*
Yin Hui,
Female, Master's student;
Research direction: Specializing in mathematics education research;
Email: 1772982786@qq.com

DOI: http://doi.org/10.12345/jetm.v9i2.24764

Wu, 2016b).

Educational assessment is a crucial yet complex component of educational activities, serving as a key mechanism for monitoring teaching effectiveness and evaluating learning outcomes, and plays a vital role in improving education quality. Test items are influenced by multiple factors, including validity, reliability and difficulty, with the last being particularly significant as it largely determines the fairness of assessments. Currently, integrated difficulty models have been explored across various fields. However, methodological limitations persist in determining the weighting between different factors within difficulty coefficient models, as well as the weighting across different levels of the same factor. Moreover, previous studies have mostly focused on the visual presentation of statistical data, often relying on subjective perception to judge whether data differences are significant, rather than conducting in-depth data analysis grounded in scientific statistical methods.

To address the limitations in current research, this study explores two innovative approaches aimed at providing new perspectives for test paper difficulty analysis. First, we introduce the AHP-based integrated difficulty coefficient model developed by Wu Xiaopeng et al. Traditional simplistic scoring methods (i.e. Natural Weighting Method) often fail to thoroughly analyze test item difficulty. By constructing expert judgment matrices to evaluate each factor and employing the AHP model to scientifically determine factor weights, this approach enables precise measurement of each factor's impact on item difficulty, thereby overcoming the shortcomings of conventional methods. Second, we integrate quantitative analysis with visualization techniques. In our examination of the 2024 NCEE Mathematics Papers (National Paper A and New Curriculum Standards I & II), we incorporated analysis of variance (ANOVA) to compare differences in difficulty coefficients and integrated difficulty coefficients across test versions. ANOVA provides an objective assessment of difficulty stability and dispersion patterns across different papers. When combined with visualization techniques that transform complex data into intuitive charts, this dual approach offers a more comprehensive and accessible reference for test difficulty research.

2. Methods

2.1 Data Selection and Analytical Methods

This study focuses on the mathematics test papers from the 2024 National College Entrance Examination (NCEE), specifically analyzing the National Paper A (NEPA), New Curriculum Standard I (NCS-I), and New Curriculum Standard II (NSC-II). The National Paper A consists of 23 test items, including two optional questions, while both New Curriculum Standard I and II contain 19 items each with no optional questions. To thoroughly examine the characteristics of these three test versions, the research employs a combination of comparative analysis and statistical analysis, aiming to provide valuable insights for educational researchers and practitioners.

2.2 Analytical Tools

2.2.1 Integrated Difficulty Model: An AHP-Based Analytical Framework

Based on the fundamental principles of AHP theory and the judgment matrices constructed by expert teams in previous literature, this study categorizes the factors of the integrated difficulty model into seven dimensions: context, parameter inclusion, operation complexity, reasoning ability, knowledge coverage, thinking orientation, and cognitive level (Zhang & Wu (2016b)). The definitions, weights, and coding schemes for each dimension are presented in Table 1.

Following the operational definitions of difficulty factors presented in Table 1, all items across the three versions were systematically coded. Subsequent statistical processing through Formula ① yielded the difficulty coefficient (d_i) for each individual factor.

$$d_i = \frac{\sum_{i=1}^{m} n_{ij} k_{ij}}{n} \left(\sum_{j=1}^{k} n_{ij} = n, i = 1,2,3,4,5,6,7\right). \quad ①$$

where n_{ij} denotes the number of test items at the j-th level of the i-th factor in each paper, represents the weight assigned to the j-th level of the i-th factor, m indicates the number of distinct levels within each difficulty factor, and is the total number of items in each examination paper. The integrated difficulty D of the entire test is calculated by Formula ② :

$$D = \sum_{i=1}^{7} d_i k_i \quad ②$$

where $k_i = (0.4, 1.20, 0.83, 2.50, 0.40, 0.83, 0.83)$ falserepresents the weight of each difficulty factor.

2.2.2 One-Way ANOVA for Completely Randomized Design

When comparing means across multiple groups, Analysis of Variance (ANOVA) is typically employed to comprehensively determine the significance of differences among several means. To examine the significance of differences between different levels of a single factor, one-way ANOVA can be applied to test for significant differences among multiple independent sample means—specifically, a completely randomized design (CRD) ANOVA.

DOI: http://doi.org/10.12345/jetm.v9i2.24764

Table 1. Classification and Weight Assignment of Levels in the Integrated Difficulty Model

Difficulty Factor	Tier	Conceptual Description	AHP Weight	Item Code
Contextual Factors	No Context	Pure mathematical knowledge as the sole context	0.27	A1
	Everyday Context	Real-life situations	0.90	A2
	Scientific Context	STEM-related contexts	1.83	A3
Parameter Factors	Parameter-free	The test item contains no parameters, requiring computation and analysis of given data only	0.50	B1
	Parameterized	The test item contains parameters that require dynamic computation, achieved through parameter processing and analysis.	1.50	B2
Operational Proficiency	Basic numerical computation	Operations involving basic numerical calculations: arithmetic, exponentiation, and root extraction	0.20	C1
	Advanced numerical computation	Operations involving complex calculations: exponential, logarithmic, and trigonometric	0.68	C2
	Basic symbolic manipulation	Operations involving trigonometric values, binomials, basic probability calculations, and related operations.	0.68	C3
	Advanced symbolic operation	Operations involving complex proofs and solving sophisticated trajectory equations and related operations.	2.44	C4
Reasoning Ability	Simple Reasoning	The reasoning content is simple and can be completed within three steps.	0.50	D1
	Complex Reasoning	The reasoning content is complex and requires more than three steps of reasoning.	1.50	D2
Knowledge Coverage	Single-concept item	The test item assesses a single key concept.	0.30	E1
	Dual-concept item	The test item assesses two key concepts.	0.78	E2
	Multi-concept item (≥3 concepts)	The test item assesses three or more key concepts.	1.92	E3
Thinking Orientation	Sequential reasoning	Using existing problem-solving approaches to address problems directly in sequential order	0.64	F1
	Reversal reasoning	Solving problems indirectly through backward utilization of established knowledge	1.36	F2
Cognitive Level	Comprehension	Comprehension of mathematical concepts, properties, and theories, relating to declarative knowledge	0.33	G1
	Application	Application of mathematical concepts, properties, and theories, relating to procedural knowledge	0.96	G2
	Analysis	Construction of appropriate mathematical models through in-depth analysis and synthesis of given conditions for problem-solving	1.71	G3

Since the sample sizes corresponding to each test paper are equal across all dimensions of difficulty factors, the F-ratio of between-group to within-group variance can be used to test whether significant differences exist among the test papers under the same contextual factors. Based on this, the following hypothesis is proposed: $H_0 : \mu_1 = \mu_2 = \mu_3$, H_1:At least two population means differ. If the test shows no significant difference ($p > 0.05$), it indicates no statistically significant differences among the test papers across the difficulty factors.

2.2.3 Analysis of Coding Examples

(2024 National Paper A, Multiple-Choice Question 6) Let the function $f(x) = \frac{e^x + 2\sin x}{1+x^2}$, then the area of the triangle formed by the tangent line to the curve $y = f(x)$ *at* $(0,1)$ and the coordinate axes is:

A.$\frac{1}{6}$ B.$\frac{1}{3}$ C.$\frac{1}{2}$ D.$\frac{2}{3}$

Difficulty Level Analysis: This problem is set in a pure mathematical context and requires students to perform routine symbolic manipulations. It involves the follow-

DOI: http://doi.org/10.12345/jetm.v9i2.24764

ing knowledge components: geometric interpretation of derivatives, calculation of tangent line equations, area computation. The solution demands only simple logical reasoning, where students directly apply learned concepts and properties through sequential reasoning. Thus, this item is coded as: A1, B2, C3, D1, E3, F1, G2.

3. Results

Applying the established analytical framework, we conducted systematic coding and statistical analysis of all three 2024 NCEE mathematics test forms, with the aggregated results displayed in Table 2.

Table 2. Difficulty Statistics by Factor Across Three Test Papers

Difficulty Factor	Tier	AHP Weight	Number of items			Percentage			Coefficients of difficulty factors		
			NEPA	NCS-I	NSC-II	NEPA	NCS-I	NSC-II	NEPA	NCS-I	NSC-II
Contextual Factors	No Context	0.27	22	17	18	95.65%	89.47%	94.74%	0.30	0.34	0.30
	Everyday Context	0.90	1	2	1	4.35%	10.53%	5.26%			
	Scientific Context	1.83	0	0	0	0.00%	0.00%	0.00%			
Parameter Factors	Parameter-free	0.50	5	5	7	21.74%	26.32%	36.84%	1.28	1.24	1.13
	Parameterized	1.50	18	14	12	78.26%	73.68%	63.16%			
Operational Proficiency	Basic numerical computation	0.20	2	2	2	8.70%	10.53%	10.53%	1.40	1.37	1.65
	Advanced numerical computation	0.68	1	1	1	4.35%	5.26%	5.26%			
	Basic symbolic manipulation	0.68	10	8	5	43.48%	42.11%	26.32%			
	Advanced symbolic operation	2.44	10	8	11	43.48%	42.11%	57.89%			
Reasoning Ability	Simple Reasoning	0.50	12	8	10	52.17%	42.11%	52.63%	0.98	1.08	0.97
	Complex Reasoning	1.50	11	11	9	47.83%	57.89%	47.37%			
Knowledge Coverage	Single-concept item	0.30	4	5	7	17.39%	26.32%	36.84%	1.04	1.19	1.26
	Dual-concept item	0.78	12	5	1	52.17%	26.32%	5.26%			
	Multi-concept item (≥3 concepts)	1.92	7	9	11	30.43%	47.37%	57.89%			
Thinking Orientation	Sequential reasoning	0.64	11	7	6	47.83%	36.84%	31.58%	1.02	1.09	1.13
	Reversal reasoning	1.36	12	12	13	52.17%	63.16%	68.42%			
Cognitive Level	Comprehension	0.33	4	3	4	17.39%	15.79%	21.05%	1.27	1.26	1.18
	Application	0.96	6	6	6	26.09%	31.58%	31.58%			
	Analysis	1.71	13	10	9	56.52%	52.63%	47.37%			

 DOI: http://doi.org/10.12345/jetm.v9i2.24764

Following the tabulated difficulty metrics (Table 2), we employed Python-based data visualization to elucidate inter-version factor difficulty trends, complemented by ANOVA to assess statistical significance.

3.1 Contextual Factors

As shown in Figure 1, all three papers primarily focus on questions testing pure mathematical knowledge, include fewer questions set in real-life contexts, and contain none based on scientific backgrounds. Among them, the NEPA included the highest number of context-free questions, totaling 22, followed by NCS-I with 17 such questions. In terms of real-life context questions, NSC-II Paper contained the most, with 2 questions, while the NEPA and NCS-I each included only 1 question of this type.

After calculations, the between-group degrees of freedom (*dfB*) was 2, and the within-group degrees of freedom (*dfW*) was 6, yielding an *F-value* of 0.016 (rounded to three decimal places, same below), with a corresponding *P-value* of 0.985. At a significance level of 0.05, the critical value obtained from the F-distribution table was 5.143. The calculated *F-value* was lower than this critical value, and the *P-value* exceeded 0.05. Based on these results, there is insufficient statistical evidence to reject the null hypothesis. This means that, given the current sample data, there is no compelling reason to conclude that there are significant differences in the mean scores among the exam papers.

Overall, the three exam papers are predominantly composed of pure mathematics context questions, while real-life and scientific context questions are relatively scarce. This suggests that current examinations place greater emphasis on testing theoretical mathematical knowledge, while to some extent neglecting the strong connections between mathematics and both real-world applications and STEM fields.

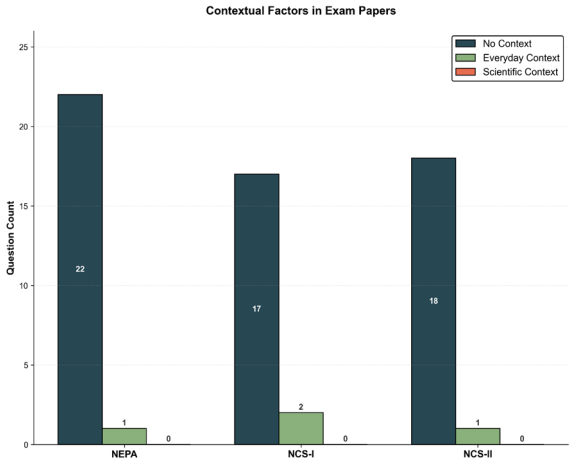

Figure 1

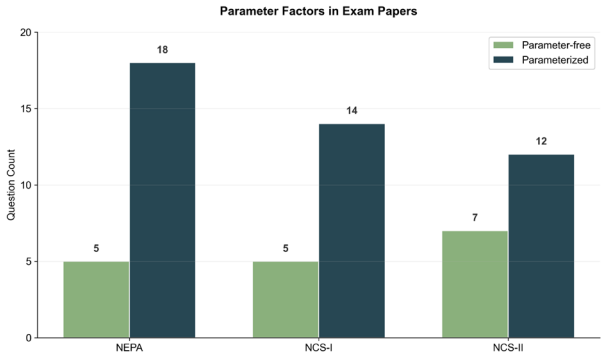

Figure 2

3.2 Parameter Factors

As observed in Figure 2, all versions primarily assess questions involving parameters. NEPA contains the highest number of parameter-based questions (18), followed by NCS-I(14), while the NSC-II has the fewest (12). Statistical analysis yields $dfB = 2$, $dfW = 3$, with $F = 0.058$ and a corresponding *p-value* of 0.945. At the 0.05 significance level, the critical *F-value* is 9.552. Since the calculated *F-value* is lower than the critical value and $p > 0.05$, the results indicate no statistically significant differences in the mean scores among the three exam papers.

Parameters serve as a bridge between variables and constants in mathematics, playing a crucial role in cultivating students' logical thinking, abstract reasoning, and problem-solving abilities. In summary, although statistical analysis indicates no significant differences in parameter-related difficulty levels across the three exam versions, their actual proportions differ slightly in terms of question quantity composition.

3.3 Operational Proficiency

As shown in Figure 3, all three exam papers place significant emphasis on assessing symbolic operations under the dimension of operational proficiency. NEPA includes 10 questions each for both basic and advanced symbolic operations. Similarly, NCS-I contains 8 questions for these two complexity levels. Notably, NSC-II has the highest number of advanced symbolic operation questions (11) among the three papers, while featuring only 5 basic symbolic operation questions. Statistical analysis yields *dfB* of 2 and *dfW* of 9, producing an *F-value* of 0.068 with a corresponding *p-value* of 0.935. At the 0.05 significance level, with a critical *F-value* of 4.256, the results show no statistically significant differences in mean scores across the exams as the calculated *F-value* is lower than the critical value and $p > 0.05$.

The substantial proportions of both basic and advanced symbolic operations in all papers demonstrates that the

DOI: http://doi.org/10.12345/jetm.v9i2.24764

education and assessment system values not only students' fundamental knowledge mastery (assessed through basic operations) but also their ability to apply this knowledge to solve sophisticated problems (evaluated via advanced operations), reflecting a comprehensive approach to mathematical competency assessment.

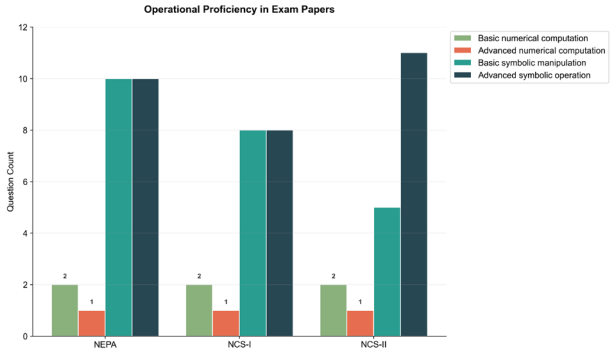

Figure 3

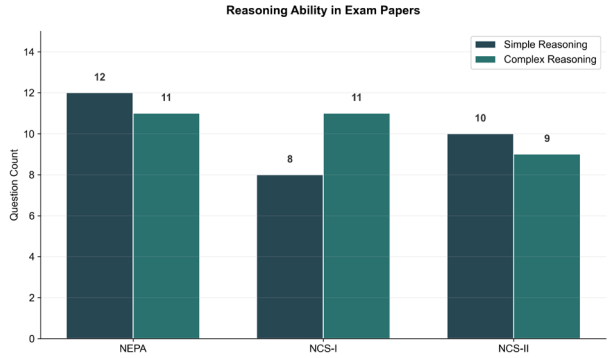

Figure 4

3.4 Reasoning Ability

The data in Figure 4 reveals characteristics in the assessment of reasoning abilities across papers: NEPA includes 12 simple reasoning questions and 11 complex reasoning questions; NCS-I contains 11 complex reasoning questions and 8 simple reasoning questions; while NSC-II comprises 10 simple reasoning questions and 9 complex reasoning questions. Statistical analysis shows *dfB* of 2 and *dfW* of 3, yielding an *F-value* of 1.455 with a corresponding *p-value* of 0.362. At the 0.05 significance level with a critical *F-value* of 9.552, the results demonstrate no statistically significant differences in the average level of reasoning ability assessment across the three papers (1.455<9.552, p>0.05). This assessment pattern reflects a balanced evaluation of candidates' basic reasoning skills and higher-order analytical abilities, while maintaining statistical equivalence in overall difficulty across different exam versions.

3.5 Knowledge Coverage

An analysis of Figure 5 reveals different emphases in knowledge coverage across papers. NEPA primarily assesses two key concepts per question (12 questions), followed by questions integrating three or more key concepts (7 questions). In contrast, NCS-I emphasizes questions combining three or more key concepts (9 questions), with fewer questions focusing on single or dual key concepts (10 questions total). NSC-II shows the strongest focus on comprehensive knowledge application, with 11 questions requiring integration of three or more key concepts, while containing only one question assessing two key concepts. Statistical analysis indicates *dfB*=2 and *dfW*=6, producing an *F-value*=0.113, and *p*=0.895. With the critical *F-value* at 5.143, the results (*F*=0.113<5.143, p>0.895) demonstrate no statistically significant differences in mean scores across the papers.

Notably, while statistically equivalent, the papers reveal different pedagogical approaches: NEPA emphasizes broad foundational knowledge through two-point questions, ensuring solid content mastery. Conversely, both New Curriculum papers prioritize deeper knowledge integration, particularly through three-or-more-point questions, encouraging students to synthesize knowledge systematically for problem-solving.

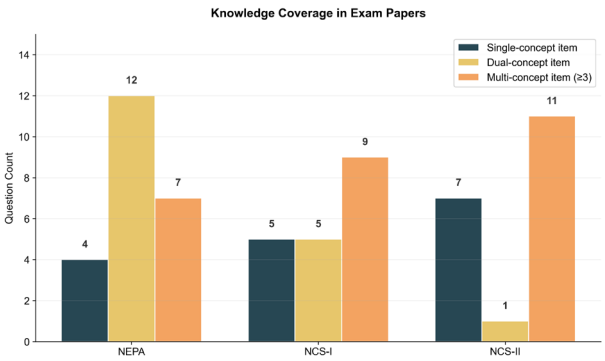

Figure 5

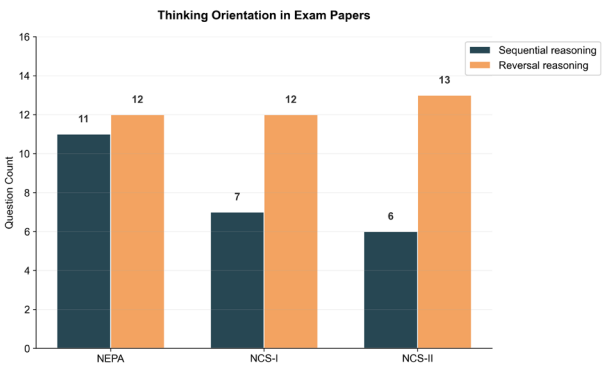

Figure 6

 DOI: http://doi.org/10.12345/jetm.v9i2.24764

3.6 Thinking Orientation

Figure 6 clearly demonstrates the variations in the "Thinking Orientation" factor across papers. Regarding sequential thinking, NEPA, NCS-I, and NSC-II show a decreasing trend, containing 11, 7, and 6 questions respectively. Conversely, for reverse thinking, these papers display an increasing trend, with question numbers of 12, 12, and 13 respectively. The statistical analysis yields *dfB*=2, *dfW*=3, *F*=0.213, and *P*=0.819. At the 0.05 significance level, with a critical *F-value* of 9.552, the calculated *F-value* is smaller than this threshold and *P*>0.05. These results provide insufficient evidence to suggest significant differences in mean scores among the papers.

Sequential thinking represents a linear, logical approach, while reverse thinking emphasizes unconventional, counter-directional reasoning abilities. Both serve as important problem-solving tools, each possessing unique value. The opposite trends observed in the proportion of sequential versus reverse thinking questions across the three papers indicate that these examinations present considerable challenges in assessing thinking abilities, requiring students to break conventional thinking patterns to solve problems.

3.7 Cognitive Level

As shown in Figure 7, the distribution of questions across cognitive levels varies among the three exam papers. At the cognitive level, NEPA, NCS-I and II contain 4, 3, and 4 questions respectively. All three papers include 6 questions each at the application level. For the analysis level, the question counts are 13, 10, and 9 respectively. Statistical analysis shows *dfB*=2, *dfW*=6, *F*=0.130, with a corresponding *P*=0.880. At the 0.05 significance level, with a critical *F-value* of 5.143, the calculated *F-value* is below this threshold (*F*=0.130<5.143) and *P*>0.05, indicating no significant differences in mean scores across the papers.

The data reveals that all three exam papers emphasize the analysis level as their primary assessment focus, demonstrating a strong emphasis on higher-order thinking skills. Analysis ability, as a core cognitive competency, encompasses problem comprehension, decomposition, and solution identification – demanding the integrated application of critical thinking, logical reasoning, and structured problem-solving skills. This consistent focus across all papers highlights the importance placed on developing students' advanced cognitive capabilities in mathematics education.

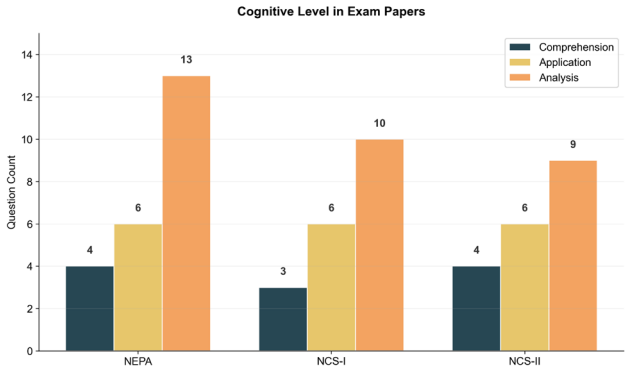

Figure 7

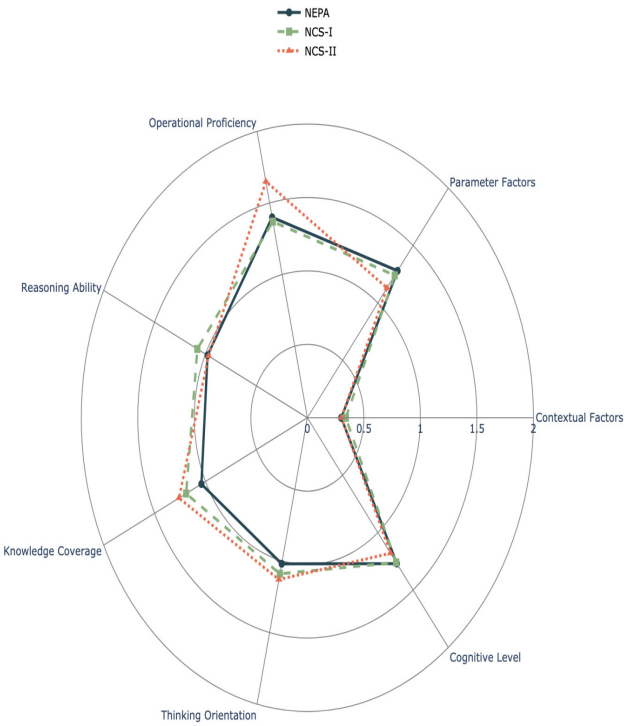

Figure 8

3.8 Integrated Difficulty Level

First, based on the data in Table 2 and the weights of various difficulty factors, the integrated difficulty levels of the three 2024 exam papers were calculated using Formula ② , with results presented in Table 3. Subsequently, a radar chart of integrated difficulty was created using the data from Table 3 (see Figure 8).

Table 3 reveals that NCS-I has the highest integrated difficulty coefficient (7.88), followed by NSC-II (7.71), and finally NEPA (7.59). Significance testing of the differences in Table 3 data yielded: *dfB*=2, *dfW*=21, *F*=0.002, with a corresponding *P*=0.998. At the 0.05 significance level, with a critical *F-value* of 3.467, the calculated

DOI: http://doi.org/10.12345/jetm.v9i2.24764

F-value is below this threshold (F=0.002<3.467) and P>0.05. These results indicate insufficient evidence to suggest significant differences in mean scores between the groups. Therefore, overall, the three exam papers maintain relatively stable integrated difficulty levels with minimal fluctuation in difficulty.

Table 3. Composite Difficulty Indices Distribution by Exam Version

	Contextual Factors	Parameter Factors	Operational Proficiency	Reasoning Ability	Knowledge Coverage	Thinking Orientation	Cognitive Level	Composite Difficulty
Weights k_i	0.40	1.20	0.83	2.50	0.40	0.83	0.83	
NEPA	0.30	1.28	1.40	0.98	1.04	1.02	1.27	7.59
NCS-I	0.34	1.24	1.37	1.08	1.19	1.09	1.26	7.88
NSC-II	0.30	1.13	1.65	0.97	1.26	1.13	1.18	7.71

Figure 8 partially reflects the differences in the composition of difficulty factors across the three exam papers. The papers exhibit quantitative consistency in the frequency of background factors, parameter factors, and cognitive levels assessed, but reveal substantial disparities in item counts across operational proficiency, reasoning ability, thinking orientation, and knowledge coverage dimensions — particularly evident in knowledge coverage and operational proficiency.

The three papers maintain near-identical assessment patterns for contextual factors, parameter factors, and cognitive levels, demonstrating commonality in basic competency requirements within the educational evaluation system. The consistency in contextual factors indicates that questions are predominantly designed within similar contexts; the uniformity in parameter factors reflects standardized requirements for difficulty control and question type distribution; while the consistent assessment of cognitive levels - a crucial indicator for measuring students' understanding and mastery of knowledge - ensures fairness and comparability of evaluation outcomes.

The observed differences in operational proficiency, reasoning ability, thinking orientation, and knowledge coverage reflect the distinct educational philosophies and teaching priorities of different exam designers, as well as their varying emphases on assessing different student competencies. The divergence in operational proficiency stems from differing computational skill requirements; reasoning ability variations manifest in the papers' emphasis on logical deduction; thinking orientation differences emerge in the distinct demands for cognitive flexibility and innovation - with the NCS-I and II Papers emphasizing divergent and critical thinking, while the NEPA focuses more on linear and normative thinking. The knowledge coverage variations are evident in the scope, depth, and breadth of key concepts assessed, where the NCS-I and II prioritize knowledge integration and application skills, whereas NEPA emphasizes in-depth exploration within specific domains. These differences necessitate targeted instructional approaches from educators.

4. Discussion

4.1 Innovative Optimization of the Integrated Difficulty Model

This study introduces an innovative integrated difficulty coefficient model based on the Analytic Hierarchy Process (AHP). By utilizing expert-constructed judgment matrices, the model achieves precise determination of weights of factors, addressing the limitations of traditional models in weight calculation. Conventional methods relying on simple additive operations or natural weighting methods often fail to accurately quantify the impact of various factors on question difficulty. In contrast, this AHP-based model, with its systematic hierarchical structure and scientific weight calculation methodology, enables more accurate decomposition of test item difficulty. This significantly enhances the precision and reliability of assessment, providing a more robust analytical tool for evaluating exam papers.

4.2 Difficulty Characteristics Revealed by ANOVA

Analysis of variance (ANOVA) indicates no statistically significant differences among the three 2024 national college entrance mathematics exam papers across key difficulty dimensions, including contextual factors, parameter factors, operational proficiency, reasoning ability, knowledge coverage, thinking orientation, and cognitive levels. In terms of overall comprehensive difficulty, the three papers also demonstrate remarkable stability with minimal fluctuations. These findings suggest a high degree of consistency in difficulty calibration across current mathematics exam papers. The ANOVA provides scientific evidence for accurately evaluating the stability and dispersion of exam difficulty, thereby strengthening the reliability of research conclusions and helping educators better understand the distribution patterns of difficulty in mathematics exams.

 DOI: http://doi.org/10.12345/jetm.v9i2.24764

4.3 Python-Powered Data Processing and Visualization

This study leverages Python's robust data processing and visualization capabilities to present relevant data from the three 2024 mathematics exam papers through intuitive charts. These visualizations clearly illustrate trends in various difficulty factors, facilitating straightforward analysis and interpretation by researchers. Additionally, Python-implemented ANOVA enables significance testing of mean differences across multiple datasets, allowing for precise identification of variations in difficulty factors among different exam papers. This approach not only provides objective data support to enhance the scientific validity and persuasiveness of conclusions but also significantly improves research efficiency. It establishes a new methodological pathway for data processing and analysis in exam difficulty research. This comprehensive approach bridges theoretical assessment models with practical educational evaluation needs, offering both methodological innovation and concrete applications for exam development.

This study adopts an innovative multi-method integrated research approach, combining AHP theory, ANOVA, and Python tools, thereby providing novel perspectives and methodologies for exam difficulty research. Future educational research should actively expand interdisciplinary and multi-method comprehensive applications, breaking free from the constraints of traditional single-method approaches. This will enhance the depth and breadth of research, yield more comprehensive and accurate findings, and drive innovative developments in educational assessment methodologies.

Given the variations among different exam papers in operational proficiency, reasoning ability, thinking orientation, and knowledge coverage - each emphasizing different aspects of student competencies - educational researchers and teachers should pay close attention to these differences. In-depth studies should be conducted on the characteristics of various exam papers to inform targeted instructional research and design. Teaching content and methods should be adjusted according to the specific competency requirements of different exams, helping students develop relevant skills and better adapt to various assessment formats. These insights can provide valuable references for educational reform and contribute to the improvement of education quality.

References

[1] Bao, J. (2002). A comparative study of the integrated difficulty of intended mathematics curricula for junior high schools in China and England. Global Education, 31(9), 48–52. (in Chinese)

[2] Liu, J., & Zhou, S. (2021). A comparative study based on the integrated difficulty model of NCEE mathematics questions: Cases of National Volume I and New NCEE Volume I. Bulletin of Mathematics, 60(4), 47–53. (in Chinese)

[3] Liu, Q., Hu, D., & Zhang, X. (2020). Analysis of NCEE question difficulty from a core literacy perspective: A case study of 2019 NCEE Mathematics National Volume (Science). Bulletin of Mathematics, 59(12), 34–40. (in Chinese)

[4] Ministry of Education of the People's Republic of China. (2020). Senior high school mathematics curriculum standards (2017 Edition, 2020 Revision) [S]. Beijing: People's Education Press. (in Chinese)

[5] Wang, J., & Bao, J. (2014). An international comparative study of the integrated difficulty of example problems in high school mathematics textbooks. Global Education, 43(8), 101–110. (in Chinese)

[6] Wu, X., & Kong, Q. (2020). Construction and application of an integrated difficulty model for NCEE mathematics questions based on AHP theory. Journal of Mathematics Education, 29(2), 29–34. (in Chinese)

[7] Yang, Z., Wu, J., & Wang, K. (2022). Analysis of National College Entrance Examination (NCEE) mathematics questions based on the integrated difficulty model: A case study of the New NCEE Volume I (2020–2022). Journal of Neijiang Normal University, 37(10), 7–12. (in Chinese)

[8] Yin, Z., Hui, Y., & Xu, F. (2024). Difficulty analysis of "trigonometric functions" questions based on the integrated difficulty model: Cases of 2021–2023 NCEE Mathematics National Volumes I, II, and Jia. Science Exam Research, 31(5), 14–19. (in Chinese)

[9] Yin, Z., Wen, S., & Xu, F. (2024). A comparative study of the integrated difficulty of NCEE mathematics questions based on AHP theory: A case study of the New NCEE Volume I (2021–2023). Research on Mathematics Teaching, 43(4), 17–21+25. (in Chinese)

[10] Zhang, Y., Wu, X., & Peng, N. (2016a). Application of the integrated difficulty coefficient model in evaluating 2016 NCEE mathematics questions. Educational Measurement and Evaluation, 12, 47–53. (in Chinese)

[11] Zhang, Y., & Wu, X. (2016b). Evaluation of New NCEE mathematics questions based on the integrated difficulty coefficient model: A case study of 2014 and 2015 NCEE Science Mathematics National Volume II. Journal of Qiannan Normal University for Nationalities, 36(3), 109–113+119. (in Chinese)

DOI: http://doi.org/10.12345/jetm.v9i2.24764