Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 9, No. 1, 2024

Optimized Design of Scoring Criteria for Large-Scale Innovation Competitions

Rui Fang[1], Bo Zhang[2], Hao Wang[3], Weihan Chen[4], Xiaoyin Wang[5, *]

[1] School of Physical Science and Technology, Tiangong University, Tianjin 300387, China
[2] School of life Sciences, Tiangong University, Tianjin 300387, China
[3] School of Economic and Management, Tiangong University, Tianjin 300387, China
[4] School of Humanities, Tiangong University, Tianjin 300387, China
[5] School of Textile Science and Engineering, Tiangong University, Tianjin 300387, China
[*] Corresponding author: wxywxq@163.com

Abstract: The reasonable and reliable evaluation score calculation scheme wins the top priority in large-scale innovation-competitions currently. Due to the disparity among individual experts, and the existing standard score calculation model merely based on the experts' own circumstance, it cannot fully reflect the comprehensive level of players, resulting in a certain degree of error in the evaluation results. From the perspective of individual and decision-making of a group, this paper improves existing models and introduces the concept of modified scores. To verify the feasibility of the model, four schemes were designed based on the analysis of data distribution. Consistency and difference factors were applied to compare. Come to a conclusion: scheme adopted the new standard score calculation model ranked the top, indicating that the new calculation model is more reliable. Next, considering that the volatile correction factor affected by data, there is a possibility of some poor actual results with high correction scores. To rectify this case, a power exponent is added to the correction factor and a sign function is adopted to specify the positive or negative of correction scores. To verify the feasibility of the model, four sets of controlled experiments were designed with the introduction of two factors: power exponent and reconsideration bonus, as well as the ranking consistency test based on the condition that the award order of the expert agreement was consistent. In the end, it was found that the scheme that introduced both power exponent and reconsideration bonus points had a sorting consistency rate of 74%, which increased by 30% compared to the original model, indicating the rationality of the modification. The expert evaluation standard score calculation model established in this paper comprehensively considers individual and group decision-making, and provides reasonable correction for the differential scoring between experts. At the same time, this scheme provides a reference basis for further optimizing large-scale innovation competition plans in the future.

Keywords: Standard score calculation, consistency test, group decision-making, large-scale innovation-competitions.

1. Introduction

In order to create a good innovation atmosphere, a large number of innovative talents who master scientific thinking, scientific methods and scientific tools have been trained and selected. A large number of innovative competitions, led by the "Challenge Cup" National Undergraduate curricular academic science and technology works by race and China International College Students' 'Internet +' Innovation and Entrepreneurship Competition, have emerged [1]. Larger innovative competitions generally adopt two-stage (online review, on-site review) or three-stage (online review, on-site review, and defense) evaluation. Due to the lack of standard answers in innovative competitions, evaluation experts need to rely on the evaluation suggestions provided by the proposer for independent evaluation, resulting in significant differences in evaluation opinions among different experts for the same entry, which in turn affects the fairness and impartiality of the evaluation results. At present, there are three main phenomena in the evaluation work of large-scale innovation competitions that affect the fairness, fairness, and scientificity of the evaluation: (1) the number of experts participating in the evaluation work is small, and the error in the evaluation work will increase; (2) There are differences in the severity of ratings among different judges during the evaluation process; (3) The lack of standard answers in innovation competitions results in vastly range of ratings for the same work by different judges.

The existing mainstream evaluation schemes mainly solve the problems caused by the above phenomena through multi-stage evaluation, standardized scoring, the use of "truncation and tail removal" method, and expert consultation. In order to maintain harmonious environment for innovation competitions and optimize more reasonable and fair schemes for competitions, Lv Shulong[2] established a scoring control model, a deviation fit model, and a differentiation model based on the final scoring data to analyze various problems caused by judges' scoring deviations; Yuan JiXue[3] defines the concept of one-time group decision-making and explores the decision-making mechanism to eliminate the subjectivity of experts in scoring; Liang Wei[4] established a comprehensive quality evaluation index system for online review and a projection pursuit model based on genetic algorithm to check the comprehensive quality of online review judges; Guo Dongwei[5] established a mathematical model for uniform distribution of papers to further reduce the impact of system errors, converted the original scores into T-scores, and used the Messi scoring method to wholly evaluate the papers; Chen Xing[6] designed a collaborative correction prediction algorithm for the synthesis of scores which passed the verification of simulation calculations.

However, existing research mostly starts with eliminating the disparity in subjective suggestions by experts and optimizing evaluation scheme. Meanwhile, the academic

level of experts varies vastly, which affects the fairness and impartiality of evaluation work. Due to the adoption of a standard score based on evaluation scheme, it is assumed that the academic level distribution of the collection of works reviewed by different experts maintain the same. However, in large-scale innovation competition reviews, only a small portion of the works reviewed by any two experts are common, and each expert can only review a fraction of the works. Therefore, the assumption of conventional standard score review scheme may not be valid. In response to the problems existing above, there is rarely existing research solutions make sense. It is urgent to explore new evaluation solutions and develop a standard score calculation model.

To effectively address the differential scoring of experts caused by different academic levels of reviewing works, this paper comprehensively considers the two-stage expert review scores, and conducts Pearson correlation analysis and T-test on the distribution characteristics of scores and the changes throughout score adjustment. Based on the test results, the effectiveness of the model has been demonstrated via adoption of statistics index such as mean and variance to establish a modified standard score calculation model and the comparison among multiple schemes. In addition, considering that the correction score of works is greatly influenced by data, there is a possibility of poor actual results with high correction scores. Therefore, a power exponent is introduced to further optimize the existing standard score calculation model, and a control experiment is designed to further verify the feasibility of the model and establish an evaluation model for the optimal evaluation plan. This provides a feasible evaluation plan for large-scale innovation competitions, which has profound significance in exploring the fairness, justice, and scientificity of the evaluation plan for large-scale innovation competitions.

2. Statistical Analysis of Data

The evaluation data of a large-scale innovation competition [7], for instance, regarding the vague distribution characteristics of the data, statistical analysis was conducted on the first expert evaluation score to study the data distribution characteristics of the initial score. Due to the fact that the scores of the first evaluation were divided into five groups, which concluded original scores (before adjustment) and standard scores (after adjustment), the distribution characteristics of the five groups and the changes throughout the evaluating process are analyzed in sequence.

2.1. Distribution characteristics of raw grades and adjusted grades

To study the overall distribution of data, most scholars typically adopt data correlation analysis methods including Pearson, Spearman, and Kendall [8-9]. Through preliminary analysis of the data (as shown in the main diagonal of Figures 1 and 2), it was found that the grading scores given by the five expert groups complies with normal distribution. Therefore, we chose Pearson correlation analysis to further study the distribution characteristics of the score data from each expert group in the first stage.

Step 1: Pearson Correlation Analysis

Due to the discreteness of the distribution of grading scores among different expert groups, the least square method can be used to fit the discrete data points with straight lines, and the slope of straight lines can be used to reflect whether there is linear consistency in the grading of different expert groups. The following figure shows the raw distribution of expert evaluation scores and the scatter plot of scoring data for different groups.

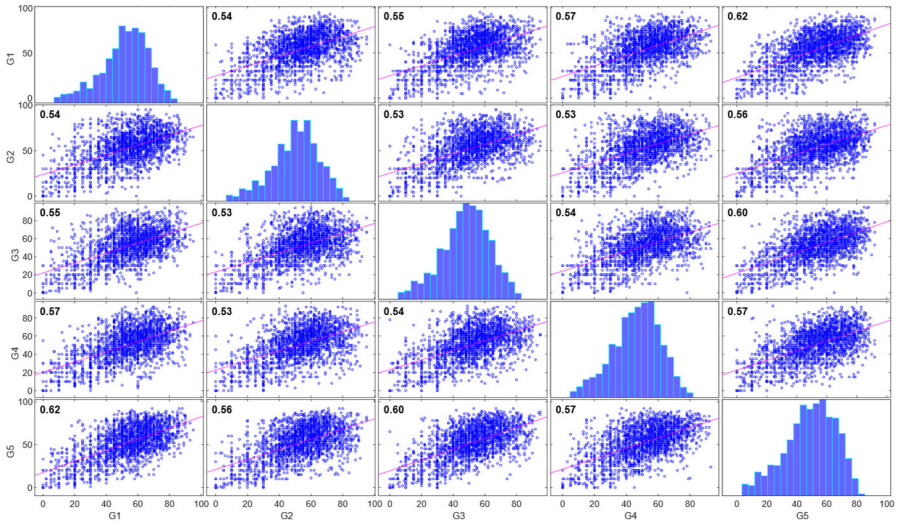


Figure 1. Rectangle distribution and scatter plot of scores before adjustment

By observing the scatter plot of the score matrix (refer to Figure 1 and Figure 2), it was found that the distribution of expert evaluation scores in each group was relatively concentrated. It can be preliminarily determined that there is a linear relationship between the five expert groups throughout the adjustment, and the distribution of scores complies with normal distribution. During the grading process, the higher linearity (correlation) of the grading scores between the five expert groups followed by the stronger impartiality of the judges' assessment. Therefore, Pearson correlation coefficient is adopted for further quantitative analysis, and the Pearson correlation coefficient calculation formula between any two samples is as follows:

$$\rho_{x,y} = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 (Y-\bar{Y})^2}}, \tag{2.1}$$

where, the size of $\rho_{x,y}$ measures the linear correlation between two variables.

By MATLAB, it was found that the Pearson correlation coefficient between the raw evaluation scores of the first five expert groups was 0.55. After adjustment, the Pearson correlation coefficient between the evaluation scores reached 0.75. Due to the correlation between the scores of two groups, it reflects the rationality of the evaluation process for different groups of experts, namely, the higher the correlation coefficient, the smaller the impact of expert group disparities on standardized reviews. Therefore, the existing sorting method based on standard scores is more reasonable compared to the original one. The following figure shows the distribution of adjusted expert evaluation scores and the scatter plot of expert scoring data for different groups.

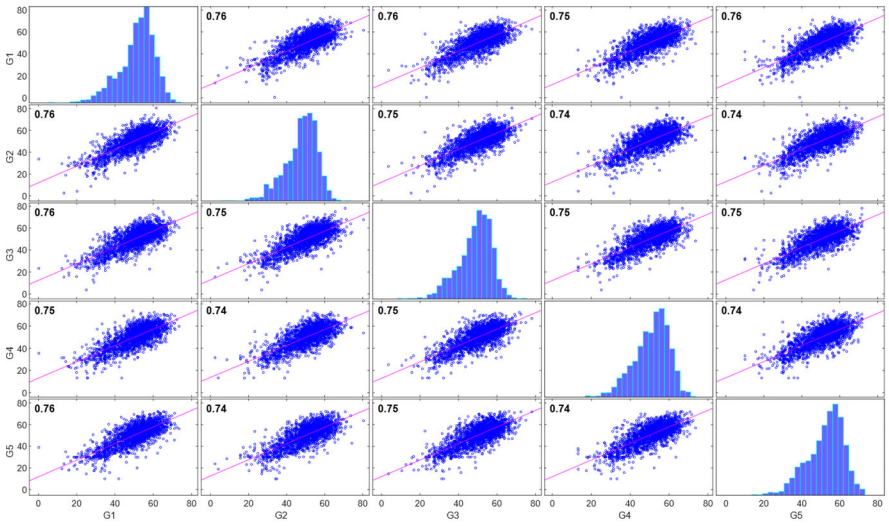
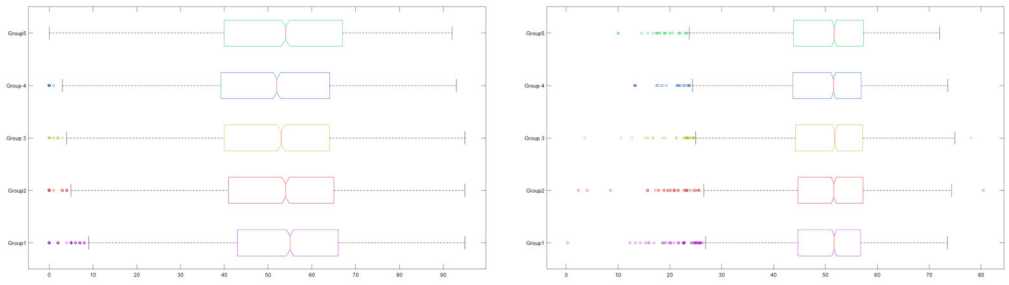


Figure 2. Rectangle distribution and scatter plot of adjusted scores

Step 2: Box Plot for Evaluation Scores

Finally, we should portrait a score box chart to further analyze the distribution of evaluation scores. The raw scores chart and the modified version are shown in the following figure.



(a) Before the adjustment (b) Aftere the adjustment

Figure 3. Box chart of scores before and after adjustment

A box plot displays the trend of the discrete distribution of data, of which the section between the upper and lower edges represents the area that is most likely to fall on. This paper adopts a box plot to statistically analyze the evaluation scores of five expert groups. By the observation of above figure, it can be seen that the mean values of the scores of the five expert groups concentrated between [50,60], and the overall data is symmetrical. However, the IQR (75% quantile to 25% quantile) of the pre adjustment scores is larger than that after adjustment, indicating that the modified scores are concentrated, and the expert scoring results are more reasonable.

2.2. Characteristics of scores changes before and after adjustment

From the above, it can be seen that there are differences in the distribution of evaluation scores before and after the adjustment. To further explore the characteristics of changes in scores before and after the adjustment, a T-test was conducted on the changes in evaluation score data of five expert groups before and after the adjustment.

Step 1: Definition of parameter variables

We set $\overline{x_a}$ the average score of the original scores of the 5 expert groups without adjustment, and $\overline{x_b}$ represent the average of the standard scores of the 5 expert groups after adjustment.

Step 2: Set the original hypothesis and alternative hypothesis

Null Hypothesis: $H_0 = 0: \overline{x_a} = \overline{x_b}$, the average score of the original one is equal to that of the standard score, which

means that adjusting the evaluation plan for the expert group's evaluation has no impact on the final result.

Alternative Hypothesis: the mean score of the original one is not equal to the mean of the standard score, which illustrates that adjusting the evaluation plan for the expert group evaluation do have an impact on the final result.

Step 3: Determine significance levels and make statistical decisions

If $\alpha \geqslant 0.05$, we accept the original hypothesis, that is, adjusting the evaluation plan for the expert group's evaluation has no impact on the final score.

If $\alpha \leqslant 0.05$, we reject the original hypothesis and accept the alternative, that is, adjusting the evaluation plan for the expert group's evaluation do have an impact on the final score.

The T-test results of 5 expert groups are shown in the Table 1.

Table 1. T-test results for original and standard scores

Expert group series index	H_0	p
Group 1	1	1.8506×10^{-11}
Group 2	1	5.0626×10^{-7}
Group 3	1	5.7603×10^{-3}
Group 4	0	0.24491
Group 5	1	1.6262×10^{-4}

According to the table 1, it can be seen that only the scores of Group 4 have remained unchanged, which have not passed the test. The other four groups of data have passed the test and verified acceptable. Accept alternative hypothesis: the adjustment and evaluation scheme reviewed by the expert group has an impact on the final result. To conclude, there is a statistical difference in scores. Due to the significant impact of expert evaluation scores on the ranking, we will conduct research on the adjusted standard score calculation model in the following text, further explore the rationality of existing formulas, and provide corresponding optimization solutions.

3. Review Scheme Design and Standard Score Calculation Model

The accuracy of standard score calculation is an important reference for defining awards in innovation competitions, therefore, an effective model that takes into account individual and group differences has become an existing crux. This paper is based on the original standard score calculation model and designs four evaluation schemes. Specifically, in the fourth evaluation scheme, the standard score calculation model is upgraded by introducing the concept of differential bonus points.

3.1. Review Plan Design

Four evaluation schemes are now put forward for competition evaluation. It is stipulated that the entries participating in the second round will always rank higher than those that did not pass the first round of evaluation. The specific design of the scheme is as follows:

Scheme 1: Truncation and Subtail Direct Accumulation Method

The evaluation will only be conducted once: for the same work, the highest and lowest scores will be removed, while the remaining original scores will be accumulated. The list of winners will be sorted based on the accumulated scores and selected in terms of proportion of awards.

Scheme 2: Truncation and Tail Removal Indirect Accumulation Method

The evaluation process is divided into two stages: firstly, we should remove the highest and lowest scores and calculate the average of remaining standard scores for the same work. The entries are sorted according to the average score, and the promotion ratio is used to select the entries for the second stage evaluation; secondly, we will sum up the calculated average score of the first stage evaluation and the standard score of the second stage, and sort them according to the accumulated score. In addition, if there is a review score in the second stage, replace the standard score with the cumulative sum of the review scores for sorting.

Scheme 3: Remove range indirect accumulation method

It is similar to Scheme 2, but in the first stage, only when the range is greater than 20, the highest and lowest scores will be removed, and the remaining standard scores will be averaged before proceeding with subsequent steps.

Specifically, the setting of the range threshold can be adjusted on its own. It is generally considered to be greater than 20, meaning that only when the threshold is exceeded, will score review or expert discussion be conducted.

Scheme 4: Average Standard Score Accumulation Method

We propose a new model for calculating standard scores: based on the existing standard scores, we introduce differential bonus points from different judges for correction.

The evaluation is divided into two steps: firstly, the original score is calculated via the new formula with the sorted average. The promoted works are selected using the promotion ratio for the second stage evaluation; Secondly, the original score is also calculated by the new model. The calculated average score of the first stage evaluation is accumulated and summed up with that of the second stage, and sorted according to the accumulated score.

3.2. Standard score calculation correction model

Due to the phenomenon of some experts' fluctuating scoring, a normalized standard score calculation formula is used to process the original scores and eliminate differences in expert ratings, thus ensures the fairness and impartiality of the award results. Assuming that the academic level distribution of the works reviewed by different reviewers is the same, here comes a standard score calculation formula:

$$x_k = 50 + 10 \times \frac{a_k - \bar{a}}{s}, \tag{3.1}$$

where, s represents the variance of the original subsample provided by a review expert, and: $\bar{a}$ represents the mean of the original subsample.

Considering that the evaluation works in large-scale innovation competitions are randomly distributed, and the degree of overlap between the papers reviewed by any two experts is limited, resulting in uneven distribution of

academic levels among works reviewed by different experts, the above assumption is not readily supported. Therefore, the following text will refine the standard score calculation model from two dimensions: horizontal comparison and vertical comparison.

Vertical comparison: Given the uneven distribution of experts' academic level, it is inappropriate merely to normalize the set of each expert, which unable to fully reflect the differences between different experts. Therefore, the scores of each expert are incorporated into the original set provided by the overall review experts, and the concept of overall correction factor is introduced:

$$\phi = \frac{a_k - \overline{a_{total}}}{s_{total}}, \tag{3.2}$$

where, s_{total} represents the variance of the original sub-samples provided by the overall evaluation experts, and $\overline{a_{total}}$ represents the mean of the original sub-samples in the overall set.

Horizontal comparison: In response to the possibility of significant fluctuations in scores, the concept of a work correction factor is introduced, defined as the variety between the scores of each judge in a single and the average score of all judges in that work. The formula is as follows:

$$\varphi = m_i - m_{mean}, \tag{3.3}$$

where, m_i represents the original score of an expert with number i for a certain piece of work, and m_{mean} represents the average original score of a review of a certain piece of work.

By integrating (3.2) and (3.3), we stipulate that the standard score calculation correction model is as follows:

$$x_k = 50 + 10 \times \frac{a_k - \bar{a}}{s} - \frac{a_k - \overline{a_{total}}}{s_{total}} \cdot (m_i - m_{mean}). \tag{5}$$

Analysis of the formula above:

For the same piece, if a judge's rating is excessively high or low compared to other judges, it needs to be corrected in the original calculation model. Therefore, the calculation model introduces a work correction factor, effectively addressing the issue of scores given by different judges exhibiting too much variance for the same work

For an individual expert, due to the uneven distribution of academic level in the reviewed works: for example, a large number of works with low academic level have highly innovative works. If only the score characteristics of this judge's reviewed works are considered in the standard score calculation, it is not conducive to the evaluation of all works. Therefore, the calculation model introduces an overarching correction factor, which uses the quantitative characteristics of the overall original score to correct its original score, thereby reducing the individual differences caused by the uneven distribution of academic levels of different expert-reviewed works.

In summary, the established standard score correction model can be well adapted to the process of large-scale innovation competitions through theoretical analysis.

4. The Evaluation Model for The Optimal Evaluation Plan

4.1. Review Results of Four Review Plans

Using MATLAB to calculate the original scoring scores in terms of the strategies of Scheme 1 to Scheme 4, and then perform a comprehensive sorting. The results are presented in the table below.

Table 2. Results of four evaluation schemes (Part)

Scheme 1		Scheme 2		Scheme 3		Scheme 4	
Serial Number	Grade	Serial Number	Grade	Serial Number	Grade	Serial Number	Grade
159	252	1	271.87	1	272.10	1	272.06
144	248	2	267.01	2	266.92	2	266.84
27	247	5	260.97	5	260.86	5	261.20
286	247	4	258.89	4	258.83	4	258.71
300	247	13	255.95	13	255.53	13	255.48
5	243	6	254.76	6	254.15	6	254.05
246	242	9	253.08	8	251.98	9	251.82
297	239	8	252.67	9	251.69	8	251.70
55	238	12	248.40	12	249.02	12	248.93
223	238	10	244.18	10	244.42	10	244.75
103	236	15	243.98	15	244.28	15	244.23
204	236	22	243.37	22	243.60	22	243.61
113	235	3	242.51	3	243.03	16	242.96
180	235	21	242.35	16	242.90	3	242.91
86	234	16	241.84	19	242.61	19	242.46
93	234	19	241.66	21	242.29	21	242.14
228	234	27	240.44	27	239.51	27	239.43
40	233	35	237.39	47	238.35	47	238.34
104	233	17	237.35	17	237.52	17	237.50
4	231	47	237.27	35	237.40	35	237.45
78	230	20	237.18	20	237.03	20	236.97
199	230	31	236.63	34	236.03	34	236.37
225	230	26	236.31	26	235.77	26	235.99
272	230	34	236.00	28	234.91	28	234.76
277	230	44	235.58	44	234.78	44	234.711

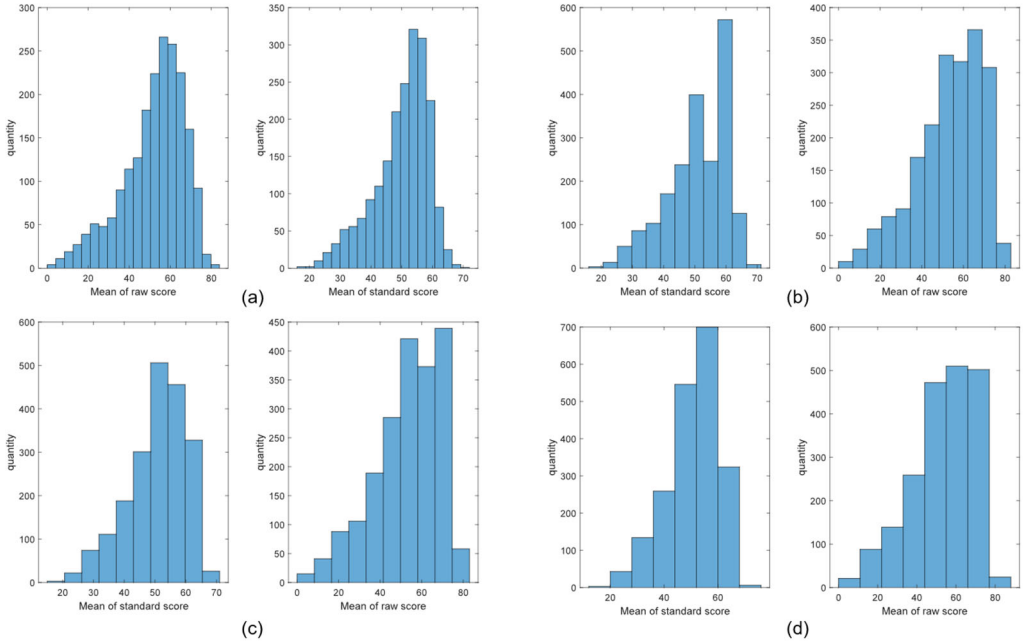


Figure 4. Distribution characteristics of scores in process of adjustment.
(a)-(d) representing schemes one to four

Result analysis:

Due to Scheme 1 only participating in the first-stage evaluation and using the original score calculation, the distribution of scores remains basically unchanged, and the overall distribution is relatively high; Both Scheme 2 and Scheme 4 are involved in two-stage evaluation, whose adjusted scores generally show a "normal distribution". The scores of the reviewed works are mainly concentrated within the range of [50,70], while the number of works with high or low segmentation is relatively small. It indicates that the adjusted score has improved compared to the original, and it is unfair to directly use the original to evaluate the work.

Quantitative analysis can demonstrate that the adjusted scores are more reasonable compared to the original scores, but it is not clear which type is better. Therefore, further indicators need to be established to evaluate the barn and born of the four types. Given that the ranking of the top 20 selected data with accuracy is determined by expert consultation, we will start from two aspects: consistency and difference test, define weighted evaluation indicators, establish the optimal evaluation model for the evaluation plan, and then evaluate.

4.2. Consistency check

Due to the fact that the ranking of the champion teams in the selected data is agreed upon through expert consultation with fixed ranking, which is used as the standard result. The ranking of the teams that have won the first prize in the four design schemes is compared with it to determine whether the theoretical calculation results of the design evaluation scheme are consistent with the standard results. Therefore, the consistency factor can be defined as follows:

$$\alpha = \frac{m_s}{27}, \tag{4.1}$$

where, $m_s$represents the number of first prizes that match the order of the s scheme and the standard results. $\alpha \in [0,1]$, The closer to 0, the better the consistency between this plan and the standard score.

4.3. Differential testing

By calculating the overall mean and variance of the final scores in different schemes and comparing them with the error rate of the standard score, a difference test is conducted to compare the pros and cons of the design schemes. Therefore, the mean error rate β is defined by comparing the overall mean a' of the final score with the overall mean a of the standard score:

$$\beta = \frac{\overline{a_{total}}' - \overline{a_{total}}}{\overline{a_{total}}}. \tag{4.2}$$

Define the error rate of the overall variance of the final grade compared to the overall variance of the standard grade:

$$\gamma = \frac{s'_{total} - s_{total}}{s_{total}}, \tag{4.3}$$

where, $\beta, \gamma \in [0,1]$, as well as the closer the error rate is to 0, the lower the difference between this scheme and the standard score.

4.4. Definition of Weighted Evaluation Indicators

The higher the consistency between the team ranking of the first prize works calculated by the design scheme and the standard ranking, the better the overall effect; The smaller the mean and variance error rate calculated by the plan, the more stable the overall plan is. Therefore, the concept of weighted evaluation indicators χ is introduced to evaluate the theoretical calculation results of the scheme, and the formula

is as follows:

$$\chi = \frac{\alpha}{\beta\gamma}. \tag{4.4}$$

After calculating the scores applied by weighted evaluation indicators χ, in order to eliminate the gaps in the evaluation scores of the four types of indicators and for better sorting and comparison, Z-score standardization is used to standardize the evaluation scores and obtain the final result. The Z-score standardization calculation formula is:

$$x_i' = \frac{x_i - \overline{x_i}}{S_i}, \tag{4.5}$$

where, x_i' is the standard post-evaluation score.

4.5. Evaluation Results of Four Schemes

We perform consistency and difference tests via MATLAB, and further calculate the scores of the four schemes by inputting the calculation results into the weighted evaluation indicators. The scores of the four schemes are shown in the table below

Table 3. Program evaluation results

Evaluation	weighted evaluation indicators χ	Mean error rate β	Variance error rate Υ	Consistency factor α
4	1.4267	0.0098853	0.06139	0.22222
3	-0.039133	0.020586	0.11199	0.18519
2	-0.68148	0.33741	0.1372	0.22222
1	-0.70609	0.33704	0.17942	0

By comparing four options: after standardization, the weighted evaluation index of evaluation option 4 is the highest, indicating that this option is the best. Further specific analysis of the results reveals that the consistency of evaluation schemes 2-4 is almost the same relying on the standard score calculation correction model; However, due to the introduction of differential bonus points between the overall correction factor and the work correction factor on the basis of the standard score calculation model, the overall mean and variance of the evaluation score in Scheme 4 are closer to the actual evaluation process, indicating that the model possesses high applicability and good calculation effect for the evaluation process of large-scale innovative competitions.

5. Model Improvement and Verification

5.1. Improvement of Standard Score Calculation Model

In the previous section on differential bonus points, we introduced a correction factor to address the evaluation differences among different judges for the same scheme. However, due to the significant influence of the data itself on the correction factor of the works, there may be situations where the score is poor but the correction factor for the works is huge, that is, the original score is far from the average value, potentially leading to an overestimation of the final standard score. Therefore, introducing exponential powers to further optimize the standard score calculation formula:

$$x_k = 50 + 10 \times \frac{a_k - \bar{a}}{s} + \frac{a_k - \overline{a_{total}}}{s_{total}} \cdot \mathrm{sgn}[(m_i - m_{mean})] \cdot (m_i - m_{mean})^{\nu}, \tag{5.1}$$

where, define power exponent as $\nu = m_i / \sum m_i$.

Specifically, a sign function is introduced to specify the positive or negative values of the correction score, and to determine whether the correction score is added or subtracted from the original one. If the original score is greater than the average score, it is necessary to perform subtraction correction on its basis and calculate the new score; Conversely, If the original score is less than the average score, it is necessary to perform addition correction on its original basis and to calculate the new standard score.

5.2. Control Experiment Design Strategy

To verify the correction effect of the standard score calculation model established in the previous text (regarded equation 5.1), considering the two influencing factors of reconsideration score and exponential power, a control experiment is designed to explore the best improvement plan. For the superiority and inferiority of experimental results, the ranking consistency rate is defined as the overlap rate between the ranking order of first prize works and the standard results calculated by the theoretical model, and then quantitative analysis is carried out on different improvement schemes.

Step 1: Set experimental objectives

With the provided evaluation result data as the standard, we should design multiple controlled experiments to explore the impact of exponential power and reconsideration score on the improved standard score calculation model, aiming to determine the final standard score calculation model

Step 2: Set evaluation indicators

We should set the evaluation indices for the standard score calculation model: sorting consistency rate, which is based on the consistent order of awards agreed by experts. Calculating the probability that the first prize ranking order is the same. The higher the probability, the better the evaluation result of the improved standard score calculation model.

Step 3: Design a controlled experiment

We should design a blank group with three control groups, and the overall control experiment table and evaluation results are shown in the table below:

Table 4. Comparison Experiment

Experimental Group	Blank Control	Control Group 1	Control Group 2	Control Group 3
Controlled Variables	Power exponent & Reconsideration score			
Dependent Variable	Ranking consistency rate P			
Inclusion factor	none	Power exponent	Reconsideration score	Power exponent & Reconsideration score
Experimental Procedure (Review option 4)	Formula (3.4) is used to calculate the standard score.	Formula (5.1) is used to calculate the standard score.	Formula (3.4) is used to calculate the standard score	Formula (5.1) is used to calculate the standard score
Result	P=0.444444444	P=0.555555556	P=0.444444444	P=0.740740741
Serial number of the first 27 works	1	1	1	1
	2	2	2	2
	5	5	5	4
	4	4	3	5
	13	13	4	3
	6	6	6	6
	9	8	9	8
	8	9	8	9
	12	12	10	10
	10	15	12	12
	15	10	11	13
	22	16	14	14
	16	17	15	15
	3	3	13	16
	19	19	22	17
	21	21	16	19
	27	47	19	20
	47	22	21	11
	17	26	20	21
	35	27	25	22
	20	28	27	23
	34	20	29	26
	26	31	23	28
	28	33	17	25
	44	34	35	27
	30	30	24	29
	31	35	34	30

*Attention: the deepening indicates that there is sequential consistency with the ordering of the standard result.

5.3. Improved Standard Score Calculation Model

According to the comparison experiment in Table 4, we can find that use of both exponential power and second stage review scores, the standard score calculation model has the best effect, with a ranking consistency rate of 74%. Compared with the blank control group: the evaluation plan without considering the influence of exponential power and second stage review scores, the ranking consistency rate has increased by 30%. Through the designed control experiment, it is quantitatively demonstrated that the improved competition evaluation plan is reasonable and the results are reliable.

In summary, the final standard score calculation model and evaluation plan design are summarized as follows:

Standard score calculation model:

$$x_k = 50 + 10 \times \frac{a_k - \bar{a}}{s} + \frac{a_k - \overline{a_{总}}}{s_{总}} \cdot \mathrm{sgn}[(m_i - m_{mean})] \cdot (m_i - m_{mean})^{\nu}. \quad (5.2)$$

Review scheme design:

The evaluation is divided into two stages: firstly, the original score is calculated by the new standard score calculation formula (see Equation 5.2), and the average is sorted. The promoted works are selected by the promotion ratio for later evaluation; Secondly, the original score will also be calculated by a new standard score calculation formula. The calculated average score from the first stage evaluation will be summed up with the standard score from the next, and sorted according to the accumulated score; In addition, if there is a review score in the second stage, we will replace the standard score with the cumulative sum of the review scores for sorting.

6. Summary and Prospect

This paper analyzes the distribution characteristics of work scores throughout the whole adjusting process by drawing scatter plots, Pearson correlation tests, and T-tests; Then, four evaluation schemes were mainly designed based on the standard score calculation formula. Specifically, the standard

score calculation formula was revised by introducing overall correction factors and work correction factors; Then, by designing consistency and difference tests and defining weighted evaluation indicators, the optimal evaluation plan is selected; Finally, comparative experiments was introduced to explore the impact of exponential powers and reconsideration scores on the final results, and a new standard score calculation model and evaluation plan were established.

Acknowledgment

This work was supported by National College Students' innovation and entrepreneurship training program of Tiangong University (202310058016,202310058030)

References

[1] Wang Xiulian. "College Students Innovative Method Competition and Innovative Talent Training". Science and Technology Entrepreneurship Monthly, 2023, 36(08):181-185.

[2] LV Shulong, Liang Feibao, Liu Wenli. "Evaluation model of judges' scores". Journal of Fuzhou University (Natural Science Edition),2010,38(03):358-362.

[3] Yuan Jixue. "Research on Group decision Mechanism in Expert Subjective rating Competition". China Soft Science, 2009 (02): 173-176+192.

[4] Liang Wei. "Comprehensive Quality evaluation of online judges based on projection pursuit Model". Statistics and Decision, 2017(23):60-63.

[5] Guo Dongwei, Ding Genhong, Liu Wei. "Optimization model of grading and ranking in essay competition". Mathematics in Practice and Understanding,2019,49(08):258-263.

[6] Chen Xing, Fang Ling, WU Songlin et al. "Collaborative modified prediction algorithm for essay score synthesis in competition". Mathematics in Practice and Understanding, 2019, 49(15): 16-23.

[7] 2023 "Huawei Cup" 20th China Graduate Student Mathematical Contest in Modeling C data, China graduate student innovation practice series competition management platform,2023.11.20.https://cpipc.acge.org.cn/cw/hp/4.

[8] Pearson, Karl. "Notes on the history of correlation", *Biometrika.* 13 (1920): 25-45.

[9] M. G. Kendall, "A new measure of rank correlation ", Biometrika, Volume 30, Issue 1-2, June 1938, Pages 81-93.